DOCUMENT RESUME

ED 370 985 TM 021 564

AUTHOR	Phillips, S. E.
TITLE	Legal Implications of High-Stakes Assessment: What States Should Know.
INSTITUTION	North Central Regional Educational Lab., Oak Brook, IL.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	RPIC-HS-93
PUB DATE	93
CONTRACT	RP91002007
NOTE	152p.; For a related document, see TM 021 563.
AVAILABLE FROM	North Central Regional Educational Laboratory, 1900 Spring Road, Suite 300, Oak Brook, IL 60521 ($19.95).
PUB TYPE	Guides - General (050) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Court Litigation; Disabilities; *Discriminatory Legislation; Educational Assessment; Educationally Disadvantaged; *Educational Planning; Elementary Secondary Education; *Legal Problems; Legal Responsibility; Policy Formation; Standards; State Legislation; *State Programs; Test Bias; Test Construction; *Test Use
IDENTIFIERS	*High Stakes Tests; *Performance Based Evaluation

ABSTRACT

States use many high-stakes assessments to make decisions about individuals. These tests may be criticized by those who believe that their purpose or application is discriminatory. Because litigation is time consuming and costly, it is advantageous for states to plan assessments carefully to maximize legal defensibility. This handbook is designed as an introduction to relevant legal issues in assessment for a variety of policymakers, although it is not intended to be a substitute for legal advice. A set of general guidelines generated from legal decisions on assessment and professional standards is presented in chapters devoted to four major areas of concern in statewide assessment: (1) testing to award diplomas; (2) potential bias against historically disadvantaged groups; (3) testing accommodations for disabled persons; and (4) legal issues in performance-assessment. Lists of terms, cases, legal theories, and measurement and educational issues introduce each chapter. Each chapter also contains recommendations for defensible policies and procedures. References, 119 in all, follow each chapter. (SLD)

****

* Reproductions supplied by EDRS are the best that can be made *

* from the original document. *

****

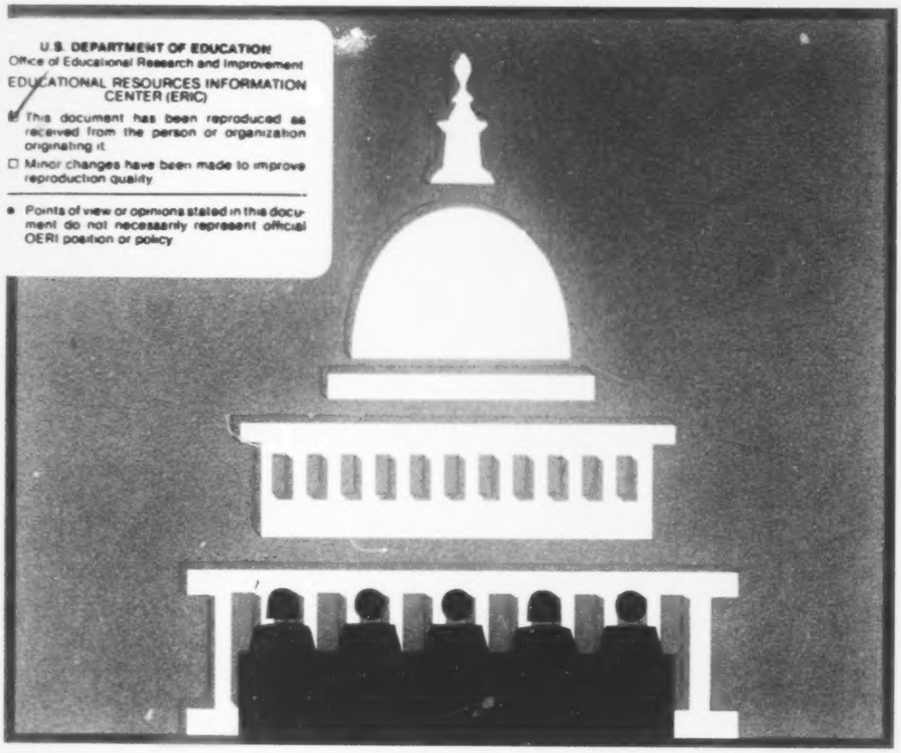
ED 370 985

SCOPE OF INTEREST NOTICE
The Eric Facility has assigned this document for proccesing to: TM

In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should refect their special points of view EA

Legal Implications of High-Stakes Assessment: What States Should Know

REGIONAL POLICY INFORMATION CENTER



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by S.E. Phillips
Michigan State University

TM021564

NCREL

BEST COPY AVAILABLE

North Central Regional Educational Laboratory
1900 Spring Road, Suite 300
Oak Brook, IL 60521
(708) 571-4700, Fax (708) 571-4716

Jeri Nowakowski:	Executive Director
Deanna H. Durrett:	Director, RPIC
Lawrence B. Friedman:	Associate Director, RPIC
Linda Ann Bond:	Director of Assessment, RPIC
John Blaser:	Editor

NCREL is one of ten federally supported educational laboratories in the country. It works with education professionals in a seven-state region to support restructuring to promote learning for all students—especially students most at risk of academic failure in rural and urban schools.

The Regional Policy Information Center (RPIC) connects research and policy by providing federal, state, and local policymakers with research-based information on such topics as educational governance, teacher education, and student assessment policy.

© 1993 North Central Regional Educational Laboratory

This publication is based on work sponsored wholly or in part by the Office of Educational Research and Improvement (OERI), Department of Education, under Contract Number RP91002007. The content of this publication does not necessarily reflect the views of OERI, the Department of Education, or any other agency of the U.S. Government.

RPIC-HS-93 ($19.95)

Legal Implications of High-Stakes Assessment:
What States Should Know

by S.E. Phillips
Michigan State University

Table of Contents

About the Author xi

Acknowledgements xiii

Executive Summary xvii
    Chapter 1—Introduction xviii
    Chapter 2—Testing to Award Diplomas xviii
        Recommendations for Developing and Implementing Legally Defensible Statewide Tests for Awarding Diplomas xix
    Chapter 3—Potential Bias Against Historically Disadvantaged Groups xxii
        Recommendations for Developing and Implementing Legally Defensible Item/Test Bias Review Procedures xxiii
    Chapter 4—Testing Accommodations for Persons with Disabilities xxiv
        Recommendations for Developing and Implementing Legally Defensible Testing Accommodations Policies xxvii
    Chapter 5—Legal Issues in Performance Assessment xxix
        Recommendations for Developing and Implementing Legally Defensible Performance Assessments xxx
    Chapter 6—Anticipated Future Legal Challenges xxxii

Chapter 1—Introduction 1
    What Is High-Stakes Assessment? 1
    Purpose of This Handbook 2
    Focus on Students and Public Education 3
    Audience 4
    Applicability of Legal Cases 4
    Terminology 4
    Glossary 5
    Disclaimer 5
    Chapter Organization 6

Chapter 2—Testing to Award Diplomas 7
    Overview 7
        Terms 7
        Case 7
        Legal Issues 7
        Measurement and Education Issues 8
        Key Questions 8
    Historical Context 8
    The *Debra P. v. Turlington* Case 9
    Legal Issues 11
        Constitutional Challenges 11
        Curricular Validity 11

Procedural Due Process 13
Vestiges of Segregation 13
Education/Validity Issues 14
Errors of Measurement 14
Curriculum/Test Match 14
Figure 1—Hypothetical Two-Dimensional Mathematics Test Matrix 16
Figure 2—Hypothetical Three-Dimensional Mathematics Test Matrix 16
Conjunctive vs. Compensatory Decisions 17
Technical Issues with Compensatory Models 18
Setting Passing Standards 20
Equating 22
Pre-equating 23
National Comparisons 24
The Lake Wobegon Effect 25
Test Security 26
Sampling Objectives 27
Equal Opportunity vs. Equal Outcome 27
The Lowest Common Denominator 28
Differentiated Diplomas 28
Naming the Test 29
Summary 30
Recommendations for Developing and Implementing Legally Defensible Statewide Tests for Awarding Diplomas 30
References 34
Cases and Statutes 34
Articles and Other Resources 34

Chapter 3—Potential Bias Against Historically Disadvantaged Groups 37
Overview 37
Terms 37
Cases 37
Legal Issues 38
Measurement/Educational Issues 38
Key Questions 38
The *Golden Rule* Lawsuit 40
Legal Issues 41
Title VII and the EEOC Uniform Guidelines 41
Equal Protection—Proving Intent 43
Fundamental Fairness Under Substantive Due Process 45
The *Golden Rule* Settlement 46
Advantages of the Settlement Terms 47
Disadvantages of the Settlement Terms 47
Effects of the *Golden Rule* Remedy on Test Validity 48
Differential Item Performance Is Not Bias Per Se 49

Test vs. Item Bias . . . 49
Identifying Biased Items . . . 50
Reliability and Validity . . . 51
Distortion of the Test Specifications . . . 53
Expert Testimony . . . 55
Reducing Group Differences . . . 56
Extension of the *Golden Rule* Remedy to Other Applications . . . 57
Summary . . . 58
Recommendations for Developing and Implementing Legally Defensible Item/Test Bias Review Procedures . . . 58
References . . . 60
Cases and Statutes . . . 60
Articles and Other Resources . . . 61

Chapter 4—Testing Accommodations for Persons with Disabilities . . . 63
Overview . . . 63
Terms . . . 63
Cases . . . 63
Legal Issues . . . 64
Measurement/Educational Issues . . . 64
Key Questions . . . 64
Physical vs. Cognitive Disabilities . . . 65
Overview of Federal Statutory Requirements . . . 67
Individuals with Disabilities Education Act (IDEA) . . . 67
Section 504 of the Rehabilitation Act . . . 68
Americans with Disabilities Act (ADA) . . . 69
Constitutional Requirements: Equal Protection . . . 69
Constitutional Requirements: Due Process . . . 70
Withholding Diplomas: The *Brookhart* Case . . . 71
State Cases . . . 72
The *Ambach* Case . . . 73
The *Hawaii* Decision . . . 74
Significance of the *Hawaii* Decision . . . 74
Measurement Issues . . . 75
Validity . . . 75
Valid vs. Invalid Accommodations . . . 76
The Purpose of Testing . . . 77
Invalid Accommodations . . . 80
Valid Accommodations . . . 81
Administrative Decision-Making . . . 82
Classification of Disabilities . . . 83
Disclosure of Accommodations . . . 84
Self-Selection with Informed Disclosure . . . 84
Summary . . . 85

Recommendations for Developing and Implementing Legally Defensible Testing Accommodations Policies 86
References 89
  Cases and Statutes 89
  Articles and Other Resources 89

Chapter 5—Legal Issues in Performance Assessment 91
Overview 91
  Terms 91
  Cases 91
  Legal Issues 92
  Measurement/Educational Issues 92
  Key Questions 92
Origins of the "Authentic Assessment" Movement 93
Historical Perspectives from Traditional Testing Cases 94
  Equal Protection 94
  Due Process 96
Legal Perspectives on Performance Assessment 96
Legal Perspectives from Employment Cases 97
  Subjective Employment Decisions 97
  Shifting Burdens of Proof 99
  The 1991 Civil Rights Act 99
Legal Perspectives from Higher Education Cases 100
  Dismissal from an Academic Training Program 100
  Revocation of a University Degree 101
Other Legal Perspectives on Performance Assessments 101
  Assigning Language Jobs Via Phone Interviews and Ethnicity 101
  Nonrenewal of Teaching Contracts 102
Measurement Issues 102
  Professional Standards 102
  Testing as a Vehicle for Curricular Reform 103
  Validity 104
  Content Sampling 106
  Scorer Reliability 109
  Standardization 112
  Test Security 112
  Potential Bias 113
  Other Technical Issues 115
Summary 115
Recommendations for Developing and Implementing Legally Defensible Performance Assessments 116
References 118
  Cases and Statutes 118
  Articles and Other Resources 119

Chapter 6—Anticipated Future Legal Challenges . . . 121
- Transition to Performance Assessments . . . 121
- Content Challenges . . . 121
- State vs. Local Control . . . 122
- Transfer Students . . . 122
- Participation in Graduation Ceremonies . . . 123
- Accommodations for Persons with Disabilities . . . 123
- Conclusion . . . 124

About the Author

Dr. S.E. Phillips is a member of the graduate faculty in the College of Education at Michigan State University, where she teaches education law and educational measurement. Specific courses include Special Education Law, Legal/Policy Issues in Testing, Testing and Grading, Standardized Testing, Measurement Theory, Item Response Theory, and Program Evaluation. She also has taught an elective course called Legal Aspects of Educational Assessment at the Thomas M. Cooley Law School in Lansing, Michigan.

Dr. Phillips's teaching in education law and her legal research on challenges to testing procedures have centered around the relationship between constitutional and federal statutory law and issues in education law, testing, and employment discrimination. Dr. Phillips has published legal articles on grade reduction penalties for high school absenteeism, diploma sanction testing, the *Golden Rule* remedy, teacher licensure testing, testing accommodations for disabled persons, performance assessment, and grade reductions for absenteeism in the *Journal of Law and Education*, the *Education Law Reporter*, and the *Cooley Law Review*.

Prior to beginning her interdisciplinary research on testing law, Dr. Phillips established herself as a national psychometric expert, with specific expertise in test scaling and equating. She has written articles on testing issues in major, nationally refereed measurement journals, including the *Journal of Educational Measurement*, *Applied Psychological Measurement*, *Applied Measurement in Education*, *Journal of Personnel Evaluation in Education*, and *Educational Measurement: Issues and Practice*. Dr. Phillips's measurement publications have addressed specific issues in test equating, quantifying equating error, curricular validity, comparison of equating methods, teacher competency, career ladder testing, test security, and discrepancy formulae and policy issues in identifying learning disabilities.

Dr. Phillips has made many presentations at national, regional, and state conferences and meetings on legal and measurement issues related to her research. She serves on the editorial board of the NCME newsletter and contributes a column on legal issues in testing. She is a member of the *Education Law Reporter* Author's Committee, she has served on the NCME External Relations Committee, and recently she served as an expert witness for a case on testing accommodations for a learning disabled applicant for teacher licensure.

Dr. Phillips's educational training includes both a Ph.D. in educational measurement and statistics from the University of Iowa and a J.D. from Thomas M. Cooley Law School. She is licensed by the State Bar of Michigan. Dr. Phillips continues to work with a variety of educational agencies on testing issues. She was asked to address a legislative conference sponsored by the Michigan Educational Assessment Program in October 1991 and recently has served on a technical advisory panel that made recommendations to the Michigan Board of Education on implementing new competency testing legislation. In her work as a consultant to the Texas, New Jersey, Mississippi, and Michigan Departments of Education; the Michigan Law Enforcement Officers Training Council; and the American Association of Medical Colleges, she has been asked to address psychometric and legal issues related to the defensibility of specific testing procedures and has had primary responsibility for equating and scaling statewide student assessment instruments. She also has served as a consultant on the development of student performance assessments to the Lansing, Littleton, and Weld County public schools.

Acknowledgements

The author wishes to thank the following professionals who reviewed earlier versions of this handbook and provided helpful guidance and suggestions. These distinguished professionals brought to the project a wide variety of experience and knowledge of the legal and measurement issues encountered in large-scale assessment programs. Professors, attorneys, directors of statewide assessment programs, and legislative staff are all represented in this group.

While the author is extremely grateful for the wise and useful advice of the reviewers, all opinions expressed in this handbook remain those of the author and should *not* be imputed to any of the reviewers or their institutions. Furthermore, the author accepts all responsibility for any errors or omissions.

The reviewers and their institutional affiliations are as follows:

Dr. Thomas Fisher, Florida Department of Education

Dr. Lee June, Michigan State University

Dr. Joan Herman, UCLA/CRESST

Dr. Elliott Johnson, National Computer Systems

Mr. William Marx, Minnesota House of Representatives

Dr. William Mehrens, Michigan State University

Mr. Kevin McDowell, Indiana Department of Education

Dr. Jason Millman, Cornell University

Dr. E. Roger Trent, Ohio Department of Education

Mr. Stanford von Mayrhauser, Educational Testing Service

The author also wishes to thank the North Central Regional Educational Laboratory (NCREL) for making this project possible, for encouragement and support during its production, and for the staff's editorial expertise in preparing the manuscript for publication. Particular thanks go to **Dr. Linda Bond**, who coordinated the project and provided helpful guidance. As with the reviewers, the opinions expressed are solely those of the author and do *not* reflect the policies or views of NCREL.

To Andre

whose future will be affected by the legal principles articulated in these pages.

Legal Implications of High-Stakes Assessment: What States Should Know

by S.E. Phillips
Michigan State University

Editor's Note: Legal Implications of High-Stakes Assessment: What States Should Know *has a very specific and important purpose: to help state and national education policymakers avoid legal challenges to their student assessment programs. With this goal in mind, Dr. Susan Phillips does an excellent job of explaining the legal and psychometric issues relevant to this need. But the North Central Regional Educational Laboratory recognizes that policymakers have other, equally important obligations—to the students who take these exams and to their parents; to the teachers whose practice is affected by the tests; to the schools, school boards, and policymakers who need quality information for educational decision-making; and to the public, which deserves to know that its investment in education is producing results. It is not enough merely to design assessments that withstand legal challenges—assessments must meet the needs of these diverse constituencies and ultimately serve education. These constituents need to know the answers to many important questions, such as the following: Are the assessments used to help students acquire important skills or are they used to isolate students into low-level remedial tracks from which they never emerge? Are the assessment results used to help all students achieve, or are they used to label some students as "good" and others as "inferior"? Are the instructional programs for all students of equivalent quality so that all students have the opportunity to perform well on the assessment? Are the assessments evaluating the knowledge and skills considered most important for students to possess? Are the best instructional methods supported by the assessments? These and other educational issues surrounding assessment policy decisions will be taken up in future assessment products. Look for* A Policymaker's Guide to the Equity Issues in Large-Scale Student Assessment *and* A Follow-Up Study of the Impact of High School Graduation Tests on Students *in 1994.*

Executive Summary

States use many high-stakes assessments, including tests to award diplomas; statewide, student, teacher, and district evaluation; licensure testing; and employment testing. These tests may be criticized by people who believe that the tests' purpose is to discriminate, even when the motives of legislators in imposing high-stakes assessment requirements are appropriate. When tests are attacked in this way, aggrieved persons, such as members of historically disadvantaged groups, which include minorities, women, and people with disabilities, may file lawsuits against the state assessment program.

Litigation is time-consuming and costly. If a judge can be persuaded to issue an injunction, the state may have to suspend assessment until the lawsuit is resolved. Thus, it is advantageous for states to plan assessment programs carefully to maximize legal defensibility.

A set of general guidelines for developing legally defensible assessment programs have been generated from legal decisions on assessment and professional standards. These guidelines cover four major areas of concern in statewide assessment: testing to award diplomas, potential bias against historically disadvantaged groups, testing accommodations for disabled persons, and performance assessment issues.

The following sections provide a brief overview of the issues covered in each chapter of the handbook, followed by a list of recommendations for legal defensibility.

Chapter 1—Introduction

High-stakes assessment refers to any assessment activity that is used for accountability. The purpose of this handbook is to provide an integrated legal and measurement overview of four specific areas of legal concern associated with high-stakes assessment programs. The focus is primarily on student assessment in public education, although some discussion of licensure and employment testing is also included. A glossary is provided at the end of the handbook for quick reference to the meanings of technical measurement and legal terms used in the handbook.

The handbook is designed to be an introduction to relevant legal issues in assessment for a variety of policymakers. Policymakers include, but are not limited to, legislators, legislative staff, lobbyists, department of education staff, state board members, testing agencies, school district boards and administrators, testing/evaluation specialists, legal advisers, and other policymakers involved in assessment enterprises. The text is written for novices who have little or no familiarity with legal or measurement concepts, but who need a basic understanding to make policy decisions regarding assessment programs.

This handbook is not intended to be a substitute for legal advice. In applying these principles to a specific set of circumstances, policymakers are advised to seek individual counsel from an appropriate legal source. Such advice may be sought from the state attorney general's office, district/corporate legal counsel, or a private attorney.

Chapter 2—Testing to Award Diplomas

In the landmark *Debra P. v. Turlington* case, African-American students who had failed a statewide test required for a diploma in Florida challenged the testing requirement as racially biased, given to affected students without adequate notice, and designed to resegregate African-American students into remedial classes. The Florida high school graduation test was a multiple-choice test of basic communication and mathematics skills applied to real life situations. In 1979, after the test had been administered three times, approximately 2% of the white seniors had not passed, compared to approximately 20% of the African-American seniors.

The *Debra P.* case established two major requirements for diploma sanction testing: adequate notice and curricular validity. Adequate notice requires that students be told what a graduation test will cover several years before the test is implemented. Curricular validity means that the schools are teaching what is being tested; under *Debra P.*, the state must collect data to demonstrate curricular validity.

Accountability has significant public support, and a variety of special interest groups are scrutinizing public education closely. When high-stakes statewide assessment programs include tests for awarding diplomas, handling potential and actual legal challenges requires careful planning, knowledge of legal and professional standards, comprehensive documentation, and a decision-making process that is procedurally fair.

Recommendations for Developing and Implementing Legally Defensible Statewide Tests for Awarding Diplomas

The following recommendations pertain generally to traditional diploma testing programs for general education students. Issues related to the newer performance assessments, item/test bias reviews, and testing accommodations for persons with disabilities are discussed in subsequent chapters. These recommendations are drawn from a variety of sources, including the information presented in the chapter, professional standards for testing, the experiences and professional judgment of the author, and common sense.

(1) Establish a technical advisory committee to advise the state agency (e.g., department of education) and state board on all policy matters and decisions related to the high-stakes assessment program.

(2) Codify all major policies in administrative rules formally adopted by the state board. At minimum, the state board should officially adopt curricular frameworks, test forms, accommodations policies, test security policies, and passing standards.

(3) If not already provided for in a state tort claims act, consider sponsoring legislation to provide limited immunity to professionals in the state who assist the state agency in the development of the assessment program.

(4) Involve representatives of major constituencies (e.g., teachers, unions, administrators, persons with disabilities, historically disadvantaged groups, business, and parents) in advisory groups providing input on assessment policies and content.

(5) Provide districts and students two to four years' advance notice of the content and format of the assessment program. Lists of specific curricular objectives, sample questions, and suggestions for appropriate test preparation provided by the state agency are helpful. Regional meetings to disseminate information and solicit input also are desirable.

(6) Provide at least as much notice to special education students and other special populations about the policies regarding assessment that will apply to them as is provided to general education students.

(7) Develop and follow a written testing accommodations policy sufficiently in advance of the first assessment date.

(8) Provide multiple opportunities for passing the test and ensure that remediation is available to those who do not pass.

(9) Document that the content being tested is being taught by the school districts in the state (curricular validity) sufficiently in advance of the date when diplomas will first be denied based on the tests, so that students have an adequate opportunity to learn the tested material. Trial administration of test forms one or more years prior to the implementation of the diploma sanction can help satisfy both the notice and curricular validity requirements.

(10) Establish passing scores as consensus standards based on a combination of content judgments and performance data.

(11) Provide a phase-in period for any new curriculum before including it on the test.

(12) Provide written materials and workshops for assisting districts in interpreting and using test score information.

(13) Design score reports that communicate effectively to those with minimal knowledge of assessment.

(14) Implement the following test security guidelines:

  (A) Ship test booklets so that they arrive only a few days before testing. Require a responsible administrator to sign a form acknowledging receipt and assuring that the materials will remain locked in a storage area with very limited access.

  (B) Allow only the minimum necessary time for testing and require all sites to test on the same day(s).

  (C) Require all testing materials to be returned immediately after testing.

  (D) Seal and number all test booklets and shrink wrap bundles of test booklets.

  (E) Require written assurance from test administrators at each site that test booklets were opened only by examinees when told to do so during testing and that no booklets were photocopied.

  (F) Require test administrators to account for all testing materials before examinees are allowed to leave the room for lunch breaks or at the conclusion of testing.

(G) Arrange for multiple proctors in each testing room and allow only one student at a time to leave during testing.

(H) Have all test administrators keep records of irregularities at the test site.

(I) Investigate all reports of breaches of test security and sanction those involved in confirmed incidents.

(J) Randomly audit test sites unannounced to ensure that proper procedures are being followed.

(K) Request the legislature to enact a statute or the state board to adopt an administrative rule defining inappropriate test preparation activities, providing sanctions for individual educators who engage in inappropriate test preparation activities or cheating, and giving the state agency authority to investigate and impose sanctions.

(L) Examine answer documents for tampering, excessive erasures, copying, and other signs of cheating. Screen group statistics and repeat testers for unusually large performance gains. Use suspicious findings to trigger appropriate investigations.

(M) Where identity may be an issue, each examinee may be required to produce photo identification, sign the answer document at the beginning of each testing session, or place a thumb print on the answer document. However, these procedures may significantly increase administration time and expense.

(15) Seek technical assistance early in the assessment program to design data collection for equating that will ensure that the achievement required to attain the passing standard remains constant from year to year.

(16) Follow professional standards in all technical matters, including, but not limited to, item development, item selection, validity, reliability, item bias review, equating, scaling, setting passing standards, test security, accommodations, test administration, scoring, and score reporting.

(17) Carefully consider the advantages and disadvantages of setting separate passing standards for each content area tested (e.g., reading, mathematics, and writing) or setting a single passing standard based on a composite total score. Involve relevant constituencies in the standard-setting process.

(18) Designate a state agency spokesperson to make all official announcements and comments about the assessment program. Caution all state employees not to make unsubstantiated statements regarding what the test measures or inferences that can be made from test scores.

(19) Provide thorough training for members of item-writing, standard-setting, content review, bias review, and scoring committees.

(20) Consult with the attorney general's office or independent counsel regarding statutory requirements and potential litigation. Detailed documentation of all actions and policies should be available for review. Such information may also be accessible by the public through freedom of information requests. Exemptions may need to be sought for secure test materials.

(21) Designate trained state agency personnel to provide continuous and comprehensive supervision and interaction with all contractors for the assessment program.

(22) Choose a neutral name for the test that does not include any constructs for which there could be debate and strong disagreement about their meaning—for example, The (state name) Graduation Test.

Chapter 3—Potential Bias Against Historically Disadvantaged Groups

Members of historically disadvantaged groups who score poorly on standardized tests have alleged that their low scores are due to bias in the test items. Specifically, they argue that test items are developed to reflect a white, middle-class culture that discriminates against persons whose culture and life experiences differ from those of the majority population.

Sometimes the potential bias seems obvious, such as when urban students are asked about farm animals that they have never seen. In other cases, the potential bias may be more subtle, such as when a vocabulary word or term is not common or has a different meaning within the historically disadvantaged group's culture. But even reviewers from historically disadvantaged groups sometimes are unable to explain precisely why one item appears biased against a particular group and another item does not. For example, similar percentages of minority students and majority students may answer the question 28 + 63 = ? correctly, while a significantly greater percentage of majority than minority students correctly answers 26 + 45 = ?

Advocates from historically disadvantaged groups also allege that the differential performance between majority and historically disadvantaged groups is a function of the different and inferior education of historically disadvantaged students. They believe that majority students have had a greater opportunity to learn the tested skills outside of formal schooling than historically disadvantaged students. Whether or not this hypothesis is correct, the question addressed in this chapter remains the same: Is the test developer precluded from measuring mastery of such skills. The larger issue of equity in assessment and instruction will be covered in a 1993-94 NCREL document.

Item bias is defined as differential performance on a test item by historically disadvantaged persons when their ability is equal to the ability of the higher performing majority group. When quantifying potential bias, some reformers have been drawn to simplistic procedures that examine differences in performance without controlling for ability. Such data formed the basis of the settlement in *Golden Rule Life Insurance Co. v. Mathias*, in which those challenging the test sought to require the test developers to choose items answered correctly by African-Americans and Caucasians in approximately equal percentages. After years of procedural battling, the case was settled out of court.

A settlement is not a court order. A settlement is an agreement between two parties to a lawsuit. In dismissing a lawsuit after settlement, the court simply acknowledges that the parties have settled their differences and no issues remain that require judicial intervention. But the court does not evaluate the content of the settlement and makes no ruling regarding it. Thus, a settlement is binding only on the parties who agreed to it and provides no legal precedent for any other lawsuit in any other court.

Despite its lack of legal authority, a settlement with one entity may be used to pressure another to agree to the same terms, which is what happened with the *Golden Rule* lawsuit. The terms agreed to in the settlement were used to pressure legislators and other testing agencies to adopt similar procedures in other contexts. For example, in *Allen v. State Board of Education*, the complainants used a ratcheted-up, more stringent version of the *Golden Rule* settlement to pressure the state into an ill-advised consent decree that severely limited the state's latitude in using a teacher certification test.

Other testing cases suggest that courts will be unsympathetic to arguments about individual items and will judge the test as a whole. Courts will be interested in expert testimony about whether the test was developed using accepted and technically defensible measurement procedures that conform to appropriate professional standards. The courts will be influenced by the views of experts from historically disadvantaged groups regarding the degree to which the test measures appropriate content and will give some deference to the state's interest in protecting the public from an undereducated citizenry. States can minimize the likelihood that a court will impose a discredited remedy by proactively seeking to implement appropriate methods for detecting and eliminating potential item and test bias.

Recommendations for Developing and Implementing Legally Defensible Item/Test Bias Review Procedures

(1) Establish a review panel of content experts representing all relevant historically disadvantaged groups (e.g., African-Americans, Hispanics, American Indians, females, and persons with disabilities) to review all items for possible offensive language, stereotypes, or cultural disadvantage prior to pretesting.

(2) Where feasible, pretest all items before use. Alternatively, scrutinize all test items for bias after-the-fact and do not score items that are judged to be unacceptable.

(3) Calculate differential item performance statistics for relevant historically disadvantaged groups using a single, professionally accepted method (e.g., item

response theory, Mantel-Haenszel). Be sure that the procedure selected compares performance for groups of equal ability.

(4) Set the criterion for flagging biased items to identify extreme outliers.

(5) Ask the review panel to re-examine all flagged items. If the panel as a whole and the members from the historically disadvantaged group for which the item was flagged believe the item is acceptable, then retain it in the item pool or score it. If not, eliminate the item from scoring, revise it and re-pretest it, or discard it and write a new item.

(6) Monitor overall test performance for each relevant historically disadvantaged group. Identify areas of weakness by group and convey this information to educators or training programs providing remediation.

(7) Disseminate outlines of the content for which examinees may be tested. Provide clear explanations and examples of item formats, test administration conditions, and score interpretation.

(8) Involve members of relevant historically disadvantaged groups at all stages of the process, including selection of content areas to be tested; development of content specifications in each selected area; making policy decisions regarding item formats, testing time, security procedures, and accommodations; forming item review and scoring panels; setting passing standards; reporting scores; and remediation.

(9) Provide expert consultation to legislators who may be pressured by lobbyists to adopt inappropriate, *Golden Rule*-type procedures.

(10) Use the news media and public relations activities to inform the public and relevant constituencies of all activities and policy decisions related to the assessment program. Enlist their cooperation by providing clear rationales for each decision, seeking their input, and answering their questions.

Chapter 4—Testing Accommodations for Persons with Disabilities

Concern for the treatment of disabled persons has become a national issue. The Americans with Disabilities Act (ADA) went into effect in 1992, requiring private entities to extend the same rights and accommodations to disabled persons as Section 504 of the Rehabilitation Act had required of public entities. Although a major provision of this legislation is to mandate the removal of physical barriers in building construction, it also prohibits discrimination against people with disabilities in employment and education.

Because the ADA was enacted only recently, case law has not yet been established under the Act. Section 504 cases suggest that the new legislation covers testing accommodations, but the courts have not indicated clearly which accommodations must be made under federal law and which may be denied.

It has been common practice to grant testing accommodations to persons with physical disabilities such as sensory deficits and mobility impairments. Because the disability was obvious to anyone who interacted with the person requesting the accommodation, verification of the disability was not necessary. Moreover, the requested accommodations were clearly appropriate, because they primarily involved the removal of physical barriers and did not significantly affect the mental skills being tested.

More recently, test administrators have received an increasing number of testing accommodation requests from persons with mental disabilities. These disabilities include attention deficit disorder, dyslexia, dysgraphia, dyscalculia, and other learning disabilities. In part, the increased number of requests may be a function of increased diagnosis and treatment in elementary and secondary schools.

Unfortunately, many of the accommodations for mental disabilities significantly affect the meaning and interpretation of the test score. Because the disability often is intertwined with the skills that the test user wishes to measure, allowing the accommodation may effectively exempt the disabled person from demonstrating the mental skills that the test measures. The test administrator then faces a policy dilemma: Should a disabled person have the option of substituting a different skill for the one measured by the test?

Federal law requires that reasonable accommodations be made for disabled persons who are "otherwise qualified." In *Southeastern Community College v. Davis*, the Supreme Court defined "otherwise qualified" as a person who, regardless of disability, can meet all educational or employment requirements. The Court held that the college was not required to modify its nursing program to exempt a profoundly hearing impaired applicant from clinical training. The Court ruled that the applicant was not otherwise qualified because she would be unable to communicate effectively with all patients, might misunderstand a doctor's verbal commands in an emergency when time was of the essence, and would not be able to function during surgery, when required surgical masks would make lip reading impossible. The *Davis* decision clearly indicates that an educational institution is not required to lower or substantially modify its standards to accommodate a disabled person, nor is it required to disregard the disability when evaluating a person's fitness for a particular educational program.

The meaning of the term "otherwise qualified" was further explained by a federal court in *Anderson v. Banks*. In this case, mentally retarded students in a Georgia school district who had not been taught the skills tested in a mandatory graduation test were denied diplomas. The court held that when the disability is extraneous to the skills tested, the person is otherwise qualified, but when the disability itself prevents the person from demonstrating the required skills, the person is not otherwise qualified. Using this definition, the *Anderson* court reasoned that the special education students who had been denied diplomas were unable to benefit from general education because of their disabilities. Because these students were not "otherwise qualified," the court reasoned, their inability to meet academic standards for receipt of a diploma should not prevent the district from establishing such standards. In the court's view, the fact that such standards had an adverse impact on disabled persons did not render the diploma test unlawful.

Another important case, *Brookhart v. Illinois State Board of Education*, held that test administrators are required under Section 504 to provide reasonable accommodations for disabled students who are otherwise qualified. The court interpreted Section 504 to require physical accommodations such as Braille or wheelchair access. However, the court stopped short of mandating all requested accommodations. The court stated that a test administrator would not be required to grant an accommodation that "substantially modified" the test. For example, the test administrator would not be required to change the test questions.

In explaining the requirement for testing accommodations, the *Brookhart* court distinguished between minimizing factors in the test format or environment that prevented a disabled person from disclosing the degree of learning actually possessed and altering the test content because a person was unable to learn the tested skills due to a disability. According to the court, a person who is unable to learn because of the disability is not otherwise qualified, and the content changes necessary for such a person to pass the test would constitute substantial modifications, which are not required by law.

In a state case related to assessment of persons with disabilities, the Hawaii Department of Education had refused a parent's request that her learning disabled son be allowed to use a reader for the statewide graduation test. The student's learning disability involved a processing deficit that substantially affected writing.

The Office for Civil Rights (OCR) agreed that allowing a reader for the reading portion of the test would defeat the purpose of the test and that denying it would not be discriminatory. But the OCR did find that denying a reader on other portions of the test, which were not designed to measure reading competency, constituted unlawful discrimination against those disabled persons who have difficulties processing written materials.

Although the OCR ruling appeared to require test administrators to provide readers for any nonreading subtest, a careful reading of the opinion suggests that the real issue in the case was due process. The OCR opinion went on to state that because the needs and abilities of disabled students vary greatly even when they have the same general disability, Section 504 requires that accommodations be judged on a case-by-case basis. However, due to a large number of requests, the Hawaii Superintendent of Education had directed staff to grant requests for readers only from blind students. Thus, OCR seemed more concerned with the procedural aspects of administrative decision-making than with predetermining the outcome of any individual testing accommodation request.

Legal and measurement analyses suggest similar conclusions regarding testing accommodations. Testing accommodations may not be denied automatically. Test administrators must evaluate each request carefully before making a decision. Requests for format accommodations should be granted if they do not change the nature of the skill being measured. Requests should not be granted if they would invalidate the inference made from the test score. Requests that fall within the grey area in between must be judged by balancing individual rights against those of the public. Whenever a test is administered under nonstandard conditions, the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices* recommend caution in interpreting test scores.

To enhance defensibility in the event of litigation, test administrators are advised to develop and disseminate written policies. They also may want to consider self-selection of accommodations with informed disclosure.

Recommendations for Developing and Implementing Legally Defensible Testing Accommodations Policies

States may choose to grant nearly all accommodation requests, to grant requests only for documented physical disabilities, or to follow a course of action somewhere in between. Whichever course a state chooses, legal defensibility will be enhanced by the development of a detailed policy and written procedures for the consideration of all requests. The policy must consider carefully both the ADA requirements and test validity, while protecting the due process rights of disabled persons. The following are suggested guidelines:

(1) Provide all school districts, training programs, and applicants for licensure with written instructions for requesting accommodations. These materials may be sent only on request, provided that their availability is communicated clearly in brochures and application materials.

(2) Provide a standardized form for requesting accommodations and clear directions for returning the application and all supporting materials to the state agency by a specified deadline.

(3) Require the requestor to provide documentation of the disability by a licensed professional experienced in diagnosing and treating the requestor's disability. A description of the disability and explanation of the necessity for the specific accommodation(s) requested should be provided in a letter signed by the licensed professional. Relevant test results and/or a description of the procedures used to make the diagnosis also might be required. The licensed professional should certify that his or her opinions are based on an in-person evaluation of the candidate conducted within the previous calendar year. In questionable cases, the licensed professional might be asked to provide documentation of his or her qualifications as an expert (e.g., a vita or biographical summary of relevant training, experience, and professional memberships, plus licenses or certifications). The requestor or licensed professional also might be asked to supply relevant medical records.

(4) Require the requestor to provide documentation of any accommodations that have been provided in the requestor's educational or training program. This documentation

should describe specific accommodations in detail and indicate the circumstances and frequency with which they were provided.

(5) If scores obtained under nonstandard conditions will be "flagged" or limited licenses granted, notify requestors of this fact and ask them to sign a statement prior to testing that confirms that they have been so notified. When the requestor is a minor, the parent(s) or guardian(s) also should sign.

(6) Designate a single individual within the state agency to review and act on all requests for testing accommodations. This person may be assisted in borderline cases by the opinion of a qualified consultant.

(7) Review testing accommodation requests on an individual, case-by-case basis, applying previously developed written criteria. Because disabilities differ in severity and an individual may have more than one disability, individual consideration is necessary. However, individuals similarly situated should be treated similarly. The state agency should develop general guidelines for accommodating various disabilities, but should review each case on its merits before making a final decision.

(8) At the state level, collect data on accommodations for mental disabilities if their effects on test validity are questionable. Such data may assist in gradually developing policies on "where to draw the line" in this area.

(9) Provide an expedited review procedure at the state level for all denied accommodation requests. Complete records of the documentation submitted by the requestor, phone calls, supporting materials received from professionals, correspondence, and the basis for the denial should be made available for the review. The review may be conducted by the agency head or a designated, qualified, impartial, outside expert hired by the agency. A written decision should be provided to the requestor.

(10) Upon written request, provide a formal appeal procedure—including a hearing—for the requestor when the denial of his or her request is upheld in the review process. Such procedures should follow the rules for administrative hearings and should allow legal representation and the presentation of evidence by the requestor. A formal hearing is useful even when not mandated by law, because it may resolve the dispute and avoid prolonged and costly litigation.

(11) Under the Individuals with Disabilities Education Act (IDEA), Section 504, and the ADA, students probably cannot be asked to bear any of the additional costs of providing testing accommodations. In licensure contexts in which the examinees bear the testing costs, reasonable additional costs for accommodations may be acceptable. Reasonable limitations of accommodations to specific testing dates and sites are probably acceptable.

(12) To ensure stability and consistency across changes in personnel, state agencies may want to codify testing accommodations policies in administrative rules or legislation. Such rules also might indicate that test proctors have the responsibility for supervising the accommodations and specify the consequences for failure to follow state agency directives regarding nonstandard testing conditions.

Chapter 5—Legal Issues in Performance Assessment

With minimal information being used to make a maximum number of high-stakes individual and group decisions, it is not surprising that critics, believing the process to be unfair, have challenged assessment programs in the courts. But what is a bit surprising is the rhetoric by advocates of performance assessments, which suggests that performance assessments can solve the problems inherent in high-stakes assessment. Instead of attacking the high-stakes uses of tests, some critics have attacked the format of the items, declaring that multiple-choice items are at fault for testing misuse. They believe that if all multiple-choice items are replaced by performance assessments, examinees will be required to demonstrate complex higher order thinking skills that are more consistent with good classroom instruction and real world applications. This claimed advantage has led advocates to refer to performance assessments as "authentic assessments." Some cognitive psychologists also believe that "authentic assessments" are superior to traditional tests because they can give greater or equal emphasis to process skills than to merely obtaining the correct answer.

However, some measurement experts doubt that performance assessments can live up to the sweeping claims made by advocates. They cite several reasons that performance assessments alone cannot solve all of our testing problems: (1) some knowledge can be measured more efficiently with objective items; (2) skilled item writers can produce challenging objective items that also measure higher order thinking skills; (3) insufficient research has been completed to document the claimed advantages of performance assessment for all testing applications; (4) performance assessments have inadequate technical properties for making high-stakes individual decisions; (5) the significantly increased costs of performance assessments are disproportionate to incremental information gains; (6) scoring of performance assessments is more subjective and thus prone to greater errors of measurement; and (7) performance assessments are more suited to classroom instruction, where incorrect decisions can be adjusted with minimal injury to the student, than to one-shot, large-scale, high-stakes accountability applications.

Inappropriate assessment practices, breaches in test security, narrowing of the curriculum, adverse impact on historically disadvantaged groups, requests for testing accommodations, measurement error, or equating problems will not magically disappear if performance assessments are substituted for traditional multiple-choice tests. Preliminary data from large-scale assessments are beginning to suggest that many of these issues in high-stakes assessment have worsened with the introduction of performance assessments. If so, legal challenges to assessment programs may increase in the future. With limited resources and tight budgets, statewide assessment programs will need to plan carefully to minimize potential litigation and to produce the documentary evidence necessary to defend high-stakes performance assessment programs in the event of a legal challenge.

Legal decisions that have addressed performance assessments have dealt with employment and higher education applications. Although these applications differ in significant ways from secondary education, they indicate the perspectives and the kinds of standards that federal courts are most likely to adopt in diploma or licensure testing challenges to performance assessments. Related cases reviewed in this chapter include challenges to subjective promotions, hiring criteria, dismissal from a training program, revocation of a college degree, the use of phone interviews and ethnic origin to assign language-related work, and nonrenewal of teaching contracts.

Prior litigation in related areas suggests that courts will apply the *Uniform Guidelines* and *Standards for Educational and Psychological Testing* to challenges of performance assessments. Although courts may be a bit more flexible in their expectations for subjective assessments, states would be well advised to proceed cautiously and to implement only those new assessments for which adequate technical data are available. This approach is particularly critical if disparate impact on historically disadvantaged groups is substantial or increases when a new performance assessment is implemented.

Adequate due process notice and appropriate validity evidence will continue to be required for performance assessments. The courts also may require evidence of scorer reliability under the fundamental fairness standard.

The subjective assessments that the courts have invalidated in the past have involved egregious procedural violations. Rarely has a court addressed the substantive and technical adequacy of a subjective procedure. However, the availability of professional standards and experts willing to testify about any flaws in an assessment program makes it unlikely that a court will assume the validity of a performance assessment without evidence. States should be cautious about making unsubstantiated claims about the advantages of performance assessments, although the need for caution should not totally preclude good faith attempts to advance the state-of-the-art.

Recommendations for Developing and Implementing Legally Defensible Performance Assessments

Designing appropriate data collection strategies during the developmental phases of an assessment program is often much easier than trying to collect the required data after the fact when a lawsuit has been filed. Knowing what is likely to be challenged and being prepared for such challenges can facilitate settlement and dissuade challengers from initiating protracted court battles. Following professionally accepted standards and carefully documenting all procedures demonstrate good faith. The following are general recommendations for increasing the legal defensibility of performance assessment programs:

(1) Follow the recommendations for diploma testing given in Chapter 2, including the test security guidelines listed under point 14.

(2) Follow the recommendations for addressing differential item performance given in Chapter 3.

(3) Follow the recommendations for developing testing accommodation policies given in Chapter 4.

(4) Provide advance notice of assessment formats and criteria for evaluating performances.

(5) Implement only those assessment procedures for which adequate data are available to document that professional standards have been met. Follow the advice of the technical advisory committee at all stages of the process of changing curricula and tests.

(6) Conduct a cost-benefit analysis to determine the areas in which performance assessment will provide important, unique information at affordable cost.

(7) Consider the potential adverse impact on historically disadvantaged groups and develop strategies for addressing the problem. Pay particular attention to potential scoring "biases" due to personal appearance, race, gender, accents, nonstandard English, poor handwriting, and so on. Use anonymous scoring whenever possible.

(8) Document opportunity to learn or job relatedness before using assessment scores to make high-stakes decisions.

(9) Administer performance tasks under standardized conditions to ensure fairness to all examinees.

(10) In addition to the usual content and bias reviews, carefully consider potential confounding of the task performance due to language deficiencies, writing or speaking deficits, personal and cultural reactions to the task, knowledge and familiarity with equipment, and other situational variables.

(11) Obtain consensus among content experts for detailed scoring criteria and train scorers to apply the criteria consistently and accurately.

(12) When feasible, obtain at least two scores for each performance and develop a procedure for identifying and resolving scorer discrepancies. When only one score is obtained for each performance, that score should be highly reliable, the total score should include multiple performances, the performances should have low weight, and/or a second score should be obtained for performances near the passing standard.

(13) Periodically and systematically recheck the ratings of each scorer for consistency and accuracy.

(14) Employ sufficient numbers of tasks and raters to ensure adequate content sampling and reliable scoring.

(15) Plan in advance for the scheduling of assessment development activities, necessary data collection, scorer training, and other contingencies so that adequate fiscal and human resources can be appropriated.

Chapter 6—Anticipated Future Legal Challenges

This chapter speculates on likely future challenges to assessment programs. Performance assessments with adverse impact on any historically disadvantaged group, test items with controversial or religious content, and denied testing accommodations for persons with disabilities are most likely to be the sources of future legal challenges to assessment programs.

Legal Implications of High-Stakes Assessment: What States Should Know

Chapter 1

Introduction

What Is High-Stakes Assessment?

High-stakes assessment refers to any assessment activity that is used for accountability. The accountability may be individual or institutional. For example, the assessment results might be used to decide which students will be awarded a diploma or state endorsement, which individuals will be granted a license to teach, which schools will receive remediation funds or merit awards, or which districts rank highest or lowest in achievement as reported by the local media. High-stakes assessment has the following general characteristics:

- Public scrutiny of individually identifiable results
- A significant gain in money, property, or prestige for those with positive assessment results
- Considerable pressure on individuals or institutions to perform well or to raise scores
- A perception that significant individual decisions are being made based on a single, imperfect piece of data over which the affected entity has *no* input or control
- Complex and costly security procedures designed to ensure maximum fairness for all who are assessed

These characteristics of high-stakes assessment suggest that a high level of anxiety is associated with the assessment and its results and that decision-making based on the assessment could potentially deprive an individual or institution of something valuable. In contrast, low-stakes assessments typically sample students and content, do not single out individuals or institutions, and do not deny or award anything of value. Hence, low-stakes assessments are less likely to be the target of a legal challenge by those who believe that the assessment program is unfair or has been used to make an incorrect decision.

Because high-stakes assessment programs are more likely to face legal challenges, this handbook concentrates on issues related to such programs. This focus on high-stakes assessments does not absolve low-stakes programs from satisfying professional standards and ensuring fairness; rather, it addresses the contexts in which policymakers are most likely to be confronted with legal challenges.

Purpose of This Handbook

The purpose of this handbook is to provide an integrated overview of four specific areas of legal concern associated with high-stakes assessment programs: (1) testing to award diplomas, (2) potential bias against historically disadvantaged groups, (3) testing accommodations for disabled persons, and (4) performance assessment.

A chapter is devoted to each of these four major topics. Each chapter describes relevant legal, measurement, and policy issues; analyzes applicable federal statutes and case law; and presents recommendations for legal defensibility. The goal is to give the reader a broad understanding of relevant legal arguments, what the courts have required in prior cases, and what one might expect from a current legal challenge. Where relevant, state decisions also are discussed.

An understanding of the legal principles involved is vital to compliance with relevant statutes and case law. The inevitable gray areas and issues that have not been fully litigated require policymakers to "read between the lines" to determine how specific legal principles might apply to their unique situations. Moreover, an understanding of the intent and underlying principles of statutes and legal decisions can help policymakers anticipate legal challenges and structure defensible assessment programs. Even if an assessment program is challenged in court, good faith attempts to follow applicable legal principles and measurement standards will assist the program in obtaining a favorable decision.

At times, the text may seem complex and difficult. But legal issues rarely lend themselves to simple interpretations, and it would be misleading to suggest straightforward solutions where none exist. The author has tried to make the text as readable as possible, to define concepts using everyday language, and to give examples. However, it is useful for the reader to become familiar with the legal terminology and complex issues that may be encountered in discussions with legal advisors or those who challenge assessment programs. A glossary is therefore provided in the back of the book. It is hoped that the detailed sections of the handbook chapters will provide useful reference material that a policymaker can consult when faced with a specific policy issue.

Statewide assessment programs face a host of issues and decisions, but the courts have specifically addressed only a handful of them. In many areas, no legal prescriptions exist. Hence, when a policymaker asks, "What does the law require?" or "What should I conclude?," the answer may be "It depends." It depends on the specific factual situation, the court in which the anticipated legal challenge will be heard, the availability of supporting expert opinion, the characteristics of the specific assessment program challenged, and so on.

To make a final decision for a particular assessment program, a policymaker must consider:

(1) The specific policy goals of the program

(2) How this decision will affect other decisions that have already been made or will be made in the near future

(3) What the policymaker feels most willing and able to defend in the particular assessment context

(4) The advice obtained from experts

(5) Relevant laws, administrative rulings, and cases in that particular jurisdiction

Even when a previous legal case addresses a particular statewide assessment issue, it will apply only in situations very similar to the facts of the decided case. When the facts of the new situation differ significantly from the original case, the court may modify its ruling consistent with the differences between the two situations.

The lack of a clear legal prescription for many of the issues discussed in this handbook may leave the reader feeling uncomfortable with the resulting ambiguity. For assessment issues that no court has decided, the only definitive legal advice is that relevant measurement standards will apply. In many legal cases related to assessment, expert witnesses provide their opinions about whether the assessment program satisfies accepted measurement practice and relevant professional standards. Most courts have given significant weight to such testimony, incorporating accepted professional standards into the legal requirements imposed on assessment programs. The author hopes that this handbook's detailed presentation of legal principles and measurement standards relevant to such issues will assist policymakers in making the "tough calls."

Finally, the recommendations at the end of each chapter are designed to summarize and highlight the major legal and measurement requirements that have emerged from the legal analysis. Some of the recommendations are practical suggestions for avoiding controversy, others come directly from statutes or cases, and still others are extensions of applicable legal principles or accepted professional practice.

Each recommendation could have been presented separately with its own discussion section. But because the recommendations are interrelated and decisions concerning a particular recommendation should not be made in isolation, the author has chosen to present the recommendations for each chapter as a unified set. The recommendations are intended to provide a starting point for discussions with a legal advisor and/or technical advisory group.

Focus on Students and Public Education

Assessment takes many forms in many contexts, including public education, higher education, licensure, and employment. However, this handbook focuses primarily on student assessment in public education. It also includes limited discussion of relevant decisions in other areas of assessment and provides some differentiation of licensure and employment contexts.

Audience

This handbook is designed to be an introduction to relevant legal issues in assessment for a variety of policymakers. This intended readership includes—but is not limited to—legislators, legislative staff, lobbyists, department of education staff, state board members, testing agencies, school district boards and administrators, testing/evaluation specialists, legal advisors, and other policymakers involved in assessment enterprises. The text is written for novices who have little or no familiarity with legal or measurement concepts but who need a basic understanding to make policy decisions regarding assessment programs.

Applicability of Legal Cases

A legal case is binding only on lower courts and administrative agencies with regulatory and adjudicatory functions in the jurisdiction in which the case was decided and applies only to cases that are factually similar. For example, a case decided in New York may be instructive to a court deciding a similar case in Indiana, but the Indiana court may choose to disregard the New York result and adopt a different legal view. If so, the two conflicting state decisions could continue to exist simultaneously in these independent state jurisdictions.

However, if the two conflicting state cases in the example above involved federal law, an appeal to a federal court could result in a legal decision consistent with either state case or different from both cases. A federal court decision applies only to state(s) within its jurisdiction. If the U.S. Supreme Court decides the case, its decision applies to all courts in the United States.

Thus, in considering the applicability of a case discussed in the handbook, the reader should carefully compare his or her own jurisdiction to the jurisdiction in which the case was decided. In many instances, legal cases may be broadly instructive but binding only in a limited area.

Terminology

When writing about sensitive issues such as discrimination, it is difficult to find terminology acceptable to everyone, in part because accepted referents have changed over time. For example, the term "African-American" has replaced "black," which replaced "Negro" as the preferred identifier. Similarly, the term "people of color" has been suggested as a preferred alternative to the term "minority." However, substituting "people of color" for "minority" has the disadvantage of excluding disadvantaged groups that are not identifiable by skin color. The gender group "females" is an example of a historically disadvantaged group that policymakers must consider but that does not fit within the term "people of color."

The legal system has used the term "historically disadvantaged group" to refer to groups that have been treated unfairly in the past. The term "historically disadvantaged" appears to describe such groups in a neutral, nonoffensive way, while broadly including disadvantaged groups identified by a variety of characteristics. However, it has the disadvantage of being a bit cumbersome to use repeatedly in text material. To indicate that all potentially disadvantaged groups are intended to be included, the author has chosen to use the term "historically disadvantaged group" in this handbook.

Except for direct quotes, the author has chosen to substitute the term "African-American" for "black." However, the term "Caucasian" has been retained to identify members of the majority group. For consistency, the term "white" may someday be replaced by the term "European-American," but at present such usage is not common.

Glossary

The handbook includes a glossary for quick reference to the meanings of technical measurement and legal terms used in the handbook. The definitions provided in the glossary are specific to the context of the handbook and may differ somewhat in other contexts.

Disclaimer

This handbook is not intended to be a substitute for legal advice. Its purpose is give a broad outline of the legal, measurement, and policy issues involved in the topics discussed. In applying these principles to a specific set of circumstances, policymakers are advised to seek individual counsel from an appropriate legal source. Such advice may be sought from the state Attorney General's office, district/corporate legal counsel, or a private attorney.

Policymakers who supervise an ongoing assessment program or who are about to implement one should seek legal advice specific to their programs for two important reasons: (1) differences in state laws and (2) recent statutory or case law changes.

The first area in which an assessment program needs individual legal advice involves interpretation of the specific state laws that govern the assessment program. This handbook primarily addresses federal law applicable to all programs and in some cases presents majority or minority views from state law. But because each of the 50 states has laws and case law precedents that are worded differently, it is important to know exactly how specific statutes and actions have been interpreted in the state with jurisdiction over the assessment program.

The second area of concern for an assessment program is timeliness. Because new laws are always being written and new cases are continually being decided, all legal publications become outdated as soon as they are published. Thus, it is important to consider the most current information available when deciding how to handle a legal matter. Individual legal counsel can provide updated information specifically tailored to the legal issue(s) faced by an assessment program.

Finally, legal advice, like any professional activity, involves experience and judgment. This handbook provides one perspective on the issues that it covers. Other facts in other places at other times may lead to different legal conclusions. Since compelling arguments often can be made on both sides of a legal issue, under certain circumstances one might want to argue for a minority view or for a change in the law. In any case, seeking individual and specific legal advice for an assessment enterprise will allow policymakers to anticipate legal challenges and make timely decisions consistent with applicable federal, state, and local laws.

Chapter Organization

The introductory material at the beginning of Chapters 2 through 5 includes a brief overview of the chapter; lists of terms, cases, legal theories, and measurement/educational issues discussed in the chapter; and a list of key questions for policymakers to consider as they read the text of the chapter. These introductory sections of each chapter are intended to function as advance organizers for the text material that follows. Each chapter concludes with a list of recommendations for implementing legally defensible policies and procedures. At the end of each chapter is a list of legal and measurement reference material that may be of interest to those who wish to explore selected issues in greater depth.

Chapter 2

Testing to Award Diplomas

Overview

The landmark *Debra P. v. Turlington* case imposed two major legal requirements on tests to award diplomas: curricular validity and adequate notice. This chapter traces the legal history of the *Debra P.* case, explains the legal standards that led the federal court to impose the curricular validity and notice requirements, and discusses key measurement and education issues related to compliance with these legal requirements.

Terms

compensatory model
conjunctive model
curricular validity
disparate impact
equating
false negatives
false positives
injunction
instructional validity
liberty interest
norms
passing standard
pre-equating
property interest
standard error of measurement
unitary schools

Case

Debra P. v. Turlington—African-American students who did not pass Florida's graduation test challenged the test.

Legal Issues

equal opportunity vs. equal outcome
Fourteenth Amendment equal protection
Fourteenth Amendment procedural due process
Fourteenth Amendment substantive due process
vestiges of segregation

Measurement and Education Issues

blaming the test
conjunctive vs. compensatory decisions
curriculum/test match
differentiated diplomas
equating and pre-equating test forms
errors of measurement
Lake Wobegon effect
naming the test
narrowing the curriculum
national comparisons
sampling objectives
setting passing standards
test security

Key Questions

(1) What is curricular validity and how does one demonstrate that a test has sufficient curricular validity?

(2) How many years' advance notice must be given to students subject to a graduation test requirement?

(3) Why did it take four years to obtain a final legal decision in the *Debra P.* case?

Historical Context

As an outgrowth of the education reform movement, many states adopted tests for awarding diplomas. These graduation tests were designed to assure parents, postsecondary institutions, and employers that all students who were awarded high school diplomas had achieved basic skills in reading, mathematics, and writing. Some states added other academic areas such as citizenship or science. The tests were intended to demonstrate that graduating students had reached a certain level of skill in these areas, giving diplomas a more consistent value. With the graduation test to back it up, a high school diploma would "mean something."

By 1988, 22 states had adopted testing requirements for awarding diplomas and several other states had statewide testing requirements that were being considered for adoption as graduation requirements. The first statewide graduation tests were adopted in the late 1970s, primarily in southern states. The list of states with diploma tests expanded through the '80s and is still growing in the '90s.

The early tests for awarding diplomas, although based on minimum-competency objectives, represented a significant departure from the local control standards that had prevailed for decades. Prior to statewide testing to award diplomas, local districts set their own graduation requirements within very broad statewide guidelines and determined which students would receive high school diplomas. Most districts did not require tests; rather, students could graduate after earning a specified minimum number of credits. Because no uniform assessment existed to ensure that all graduates could read, write, and compute at a satisfactory level, many critics charged that American schools were graduating illiterates with little chance of advancing beyond low-level, menial jobs. Some critics also charged that students were being passed from grade to grade based on "seat time" rather than achievement of academic objectives.

The earliest tests for awarding diplomas comprised multiple-choice items from which the student was asked to select the "correct" or "best" answer. Reading tests included passages of text followed by comprehension questions. Math tests used computation and story problems to assess the basic operations of addition, subtraction, multiplication, and division. Writing tests asked grammar questions and sometimes required students to write a short essay on a specified topic.

The intent of the southern legislatures that first adopted these minimum-competency graduation tests was to upgrade their respective statewide education systems so that their states could be more competitive in attracting businesses. Unfortunately, the large urban school districts in many of these states had become integrated only recently, and students from historically disadvantaged groups still felt the effects of past segregation.

Not surprisingly, when the graduation tests were first administered statewide, students from historically disadvantaged groups failed in large numbers. Although high numbers of majority students also failed, the disparity in passing rates between majority students and students from historically disadvantaged groups was substantial. Critics charged that historically disadvantaged students were being discriminated against for their prior, inferior, segregated educations and for being part of a culture that was different from the majority culture. Some evidence also suggested that even in integrated districts many historically disadvantaged students in urban areas continued to be assigned to schools with inadequate facilities and poor teachers, attended predominately by historically disadvantaged students. So even though the motives of legislators in imposing the graduation testing requirements were laudable, the tests themselves became the target of those who believed that the real purpose was to discriminate against historically disadvantaged groups.

One of the early tests that received such criticism was the Florida functional literacy examination. The ensuing class action lawsuit, *Debra P. v. Turlington*, commonly referred to as the *Debra P.* case, became the landmark case establishing legal standards for diploma testing.

The *Debra P. v. Turlington* Case

The Florida legislature passed graduation test legislation in 1976, establishing the Functional Literacy Examination (FLE) as the state's graduation test, effective for the 1979 graduating class. The FLE was a multiple-choice test of basic communication and mathematics skills applied to real life situations. Students who failed the FLE were allowed to retake it. At graduation, students who had not passed the test received a certificate of completion rather than a high school diploma.

In 1979, after three administrations of the FLE, approximately 2% of the Caucasian seniors had not passed, while approximately 20% of the African-American seniors had not passed. The *Debra P.* case was brought by several African-American students on behalf of all African-Americans who had been denied high school diplomas because they had failed the FLE. The students who challenged the FLE alleged that the test was racially biased, given to affected students without adequate notice, and designed to resegregate African-American students into remedial classes. The lawsuit alleged violations of Title VI of the Civil Rights Act of 1964 and the due process and equal protection clauses of the Fourteenth Amendment to the U.S. Constitution.

The federal district court that heard the initial case found no present intent to discriminate, but held that the FLE perpetuated the effects of past discrimination in Florida schools in violation of the statutory and constitutional provisions cited. The court further held that the notice period of approximately one year was inadequate, because the skills tested covered multiple years and classes of instruction. To remedy the due process and vestiges of segregation violations, the court enjoined the state of Florida from imposing passage of the test as a graduation requirement until 1983. At that time, all graduating seniors would have completed a full 12 years of schooling in unitary integrated schools and would have had four years' notice to prepare for the test.

The federal appeals court was asked to review the decision of the district court in the *Debra P.* case. The appeals court affirmed the district court's injunction prohibiting the withholding of diplomas until 1983, but found that the district court had not adequately addressed two issues critical to the legality of the Florida FLE testing program. The case was remanded to the district court for another trial to determine (1) whether the vestiges of segregation were still adversely affecting the achievement of African-American students and (2) whether the test had "curricular validity"—i.e., measured skills being taught in Florida schools.

At the second trial, the state of Florida was successful in demonstrating the curricular validity of the FLE based on a study directed by a private consulting firm and carried out by the department of education and local school districts. The study surveyed teachers, curricular materials, district personnel, and students to determine whether the specific skills tested by the FLE were being taught in Florida classrooms. The study concluded that, on average, Florida seniors had received 2.7 mastery-level lessons on each objective tested, although a single mastery-level lesson would have been sufficient to establish the curricular validity of each tested objective. The district court found that Florida had met its burden of proving that the FLE had curricular validity, even though the study was cross-sectional (a sample of instruction across grades at a single point in time) rather than longitudinal (following a single group of students across 12 years) and did not document the instruction for every student in every classroom.

Furthermore, the court held that even if the disproportionate African-American failure rates were caused by past discrimination, the FLE was a fair test of what was being taught and a necessary remedy. That is, the test was seen as part of the *solution* to inadequate African-American achievement, not the cause. The court did not find a constitutional violation, even though some students had mediocre teachers, and held that requiring the state to document the instruction for every student in every classroom would be an impossible burden. According to the court, the appropriate curricular validity standard for a test used to award diplomas was that "the skills be included in the official curriculum and that the majority of the teachers recognize them as being something they should teach." (p. 186) The district court opinion from the second trial was affirmed on appeal, and Florida began awarding diplomas only to students who passed the FLE.

The following sections provide a more detailed review of the legal, measurement, and education issues raised in the *Debra P.* case. The final section of this chapter summarizes recommendations for developing and administering a legally defensible diploma testing program, based on the *Debra P.* case and related measurement principles.

Legal Issues

In the *Debra P.* case, the federal courts set legal standards for graduation tests by addressing three major legal issues: (1) curricular/instructional test validity as a substantive due process requirement; (2) prior notice required to satisfy procedural due process; and (3) vestiges of segregation as intent to discriminate under the equal protection clause. Before discussing these issues, it is necessary to review briefly the threshold requirements for Fourteenth Amendment constitutional challenges.

Constitutional Challenges

A constitutional challenge under the substantive due process, procedural due process, or equal protection clauses of the Fourteenth Amendment requires state action. In a legal context, state action refers to a deprivation of constitutional rights by a state government entity, which includes state agencies, local governments, and public school districts. The Fourteenth Amendment protects against injuries to life, liberty, or property. The *Debra P.* court held that a high school diploma is a property interest subject to Fourteenth Amendment protections. The applicable section of the Fourteenth Amendment states as follows: "No state shall make or enforce any law which shall . . . deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws."

Curricular Validity

Requiring curricular validity in a test used to award high school diplomas derives from the substantive due process standard of fundamental fairness. Under the Supreme Court's rationality standard, which applies in all cases, state action must serve a legitimate government interest and must use means that are not arbitrary or capricious. Although the

Debra P. court recognized a legitimate government interest in ensuring a minimal level of competence for all students receiving diplomas, it also held that fundamental fairness required that the state demonstrate curricular validity for all tests used to further that interest.

The legal requirement that tests be valid has been uniformly applied by the courts, and it is consistent with the views of measurement experts and the *Standards for Educational and Psychological Testing* established by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). The questions to be answered in specific cases have been which type of validity is required and what data is necessary to establish validity.

The general definition of validity states that a test must measure what it is intended to measure. Validity depends on how the test score will be used; it is not a property of the test per se. In a specific testing application, some inferences from test scores may be valid, while others are not.

For descriptive purposes, measurement specialists generally classify validity evidence into four major categories: content, curricular, criterion, and, construct. Content validity refers to the relationship between the test items and the knowledge, skills, and/or abilities considered important in the domain that the test is designed to sample. Content validity also may include a curricular validity requirement. Curricular validity refers in general to the match between instruction and what is tested. Criterion validity refers to the correlation between test scores and some other variable that the test is designed to predict. Construct validity refers to experimental evidence that a test provides meaningful information about a postulated psychological trait.

Curricular validity is divided into two types: (1) curricular validity—the match between test content and curricular materials, such as textbooks; and (2) instructional validity—the match between test content and what is actually taught in the classroom. These terms are not interchangeable: If a teacher's classroom instruction differs from the published curriculum, for example, a test that has curricular validity might not have instructional validity.

However, these terms have been used somewhat interchangeably in legal cases and the term curricular validity in the legal sense appears to mean reasonable evidence that the test content is included in written curriculum materials and is being taught in most classrooms. Thus, the legal requirement for curricular validity appears to be a hybrid between traditional definitions of curricular and instructional validity. As the opinion in the *Debra P.* case indicated, a pure instructional validity standard would be unattainable, because it would be impossible to show that every student in every classroom had been taught all of the skills tested. The less stringent curricular validity standard adopted by the *Debra P.* court was a reasonable compromise incorporating elements of the match of the test content to both written materials and teacher judgments.

Procedural Due Process

The purpose of procedural due process requirements is to guarantee that a process that may deprive a person of a property or liberty interest is fair. In awarding diplomas, the major concern is to give students adequate notice to prepare for the testing requirement. However, the interests of the student, which call for early notice, must be balanced against the interests of the state, which call for upgrading the educational system as quickly and efficiently as possible.

The *Debra P.* court found that one year was not adequate notice and suggested that four to six years' notice would be more appropriate. Although the court did not mandate a specific notice period, the injunction in the *Debra P.* case did mandate a four-year notice period. Because this notice period corresponded to the time necessary for all students to have spent all 12 years of their education in nonsegregated schools, it is not clear that the court would require a full four years' notice for other graduation tests.

Vestiges of Segregation

The vestiges of segregation argument held that African-American high school students in Florida continued to be disadvantaged by the segregated schools—mandated by state law—that they had attended during their elementary years. The purpose of this argument was to establish an intent to discriminate, which is necessary for an equal protection challenge. Since the courts had already recognized segregation as intentional discrimination and therefore a violation of equal protection guarantees, then withholding diplomas based on test failure that resulted from such segregation also violated equal protection.

The equal protection clause of the Fourteenth Amendment guarantees that the government will give equal treatment to persons situated similarly. In deciding whether the government has violated equal protection guarantees, the court can use a stringent or a lenient standard of judicial review. If the challenger can convince the court that a stringent standard of review is appropriate, then it is more likely that the court will find the challenged government action unconstitutional.

The requirements for a stringent standard of review include the following: (1) two classes of persons are being treated differently; (2) one of the classes constitutes a protected racial or ethnic group (e.g., African-Americans); (3) decision-making procedures deny a property right to members of the protected group significantly more often than to members of the nonprotected group; and (4) the government's actions reveal an intent to discriminate. In the context of graduation tests, an equal protection violation would occur if African-Americans' lower test scores resulted in the government's denying diplomas to a substantially larger percentage of African-Americans than Caucasians and if evidence indicated that the government had adopted the testing requirement with the intent to cause this differential result.

Although in other contexts discriminatory *effect* is emphasized over discriminatory *intent*, in equal protection challenges the Supreme Court has retained the intent requirement. The *Debra P.* case had all of the elements required for an equal protection violation *except* the intent to discriminate. By the time of the second appeal, all African-American students subject to the testing requirement had attended integrated schools for their entire 12 years of education. Thus, the court found that although students might feel some lingering, secondary effects of segregated schools, such as supplying less qualified teachers or fewer curricula materials to predominantly African-American schools, these effects were not sufficient to establish an intent to discriminate against African-Americans. On the contrary, the court held that the testing requirement was a permissible remedy for the dual school system's past failure to teach basic skills to African-American students.

Education/Validity Issues

The criticisms of diploma testing that led to the *Debra P.* challenge included education and measurement issues as well as legal issues. In its ruling, the court addressed the validity issue in particular and also dealt with some of the educational concerns surrounding the testing requirement. The following sections elaborate on these concerns.

Errors of Measurement

Any data used to make a high-stakes decision must be carefully evaluated for accuracy. Every time a decision is made, two kinds of errors are possible. The decision-maker may erroneously grant a diploma to a student who has not really learned the specified content and skills or deny a diploma to a student who has actually learned the specified content and skills.

If the decision-maker adopts procedures that decrease one of these types of errors, the other type of error will increase. For example, if the testing requirement is abolished to avoid erroneously denying diplomas to deserving students, the number of incompetent students receiving diplomas will increase. Similarly, if one adopts a very high passing score on the test to avoid granting diplomas to any student who has not clearly learned all of the specified content and skills, some students who have adequate knowledge and skills will be denied diplomas. Neither extreme is likely to provide a satisfactory basis for awarding diplomas or to advance the state's interest in ensuring that those students who receive high school diplomas have learned the specified content and skills. Therefore, assessment requirements and passing standards must reflect a compromise between these two extremes and hold reasonable expectations for all students.

Curriculum/Test Match

Curricular validity usually is measured by quantifying the relationship between test content and the content of curricular materials or observed lessons. This quantification can be accomplished by classifying both the test and the instruction or curricular materials on the same content taxonomy matrix. Teacher surveys indicating whether the skills required for specific test items have been taught and providing a rating of the importance of those skills in the curriculum also can provide useful information for judging curricular validity.

In addition to determining the particular type of curricular validity evidence to be collected, test users must also be concerned with (1) the closeness of the match between the curricula/instruction and the test; and (2) the instructional time period over which the match should be measured. The resolution of both of these issues can have a major impact on the degree of curricular validity attributed to the test.

The closeness of the match between the instruction and the test items determines the level of student performance that can be measured. If the examples used in instruction are identical to the items on the test, it is probable that some students will obtain correct answers by regurgitating what was presented in class. This type of rote feedback provides no information regarding a student's understanding of concepts or ability to apply them. On the other hand, when test exercises involve skills covered in instruction but contain novel particulars, inferences can be made about a student's ability to solve problems and to apply learned material to new situations.

For instance, suppose that in a math class students are given the following definition and examples:

prime number: whole number evenly divisible only by 1 and itself

examples: 2, 3, 5, 7, 11, 13, 17, 19

Counter-examples:	4, 6, 8 (all divisible by 2)
	9, 12, 18 (all divisible by 3)
	10, 15, 20 (all divisible by 5)

Now consider items for testing whether students understand the concept of "prime number." If the item uses particulars that were presented in the instruction (e.g., 7, 13, 15, 18), students might answer correctly because they remembered (or memorized) what the teacher had said and not because they really understood the concept of "prime number." Unfamiliar examples (e.g., 16, 29, 33, 37) are necessary to test whether students can apply the definition of "prime number" that they have learned.

Even when test items use novel particulars, the judged overlap between the test and the instruction will be affected by the way in which the overlap is measured. One common way of measuring overlap is to develop a content matrix describing the domain of knowledge and skills on which the instruction and the test are based. Such a matrix may be two-dimensional, with process skills on one dimension and content subareas on the other dimension. The content matrix is used to map separately the content of the instruction and the content of the test items. Overlap between the instruction and the test is then quantified by the proportion of cells that they share.

For example, a math test may cover multiplication and division of fractions and decimals. The content matrix for this test, illustrated in Figure 1, would have four cells: multiplication of fractions, division of fractions, multiplication of decimals, and division of decimals. If the instruction covered only multiplication of fractions and decimals (two of the four cells) and the test items were evenly distributed across all four cells, then the overlap would be 50%. A third dimension could be added to this matrix by dividing the content of each cell into computational exercises and story problems. This new matrix, illustrated in Figure 2, would then have twice as many (2 x 4 = 8) cells in which instructional content and test content could be classified. If the instruction covered multiplication but not division or story problems (cells one and three) and the test items were evenly distributed across all eight cells, the overlap would be 25%.

Figure 1

Hypothetical Two-Dimensional Mathematics Test Matrix

Content	Operation	
	Multiplication	Division
Fractions	Cell 1	Cell 3
Decimals	Cell 2	Cell 4

Shaded areas indicate instructional coverage.

Figure 2

Hypothetical Three-Dimensional Mathematics Test Matrix

Content		Operation	
		Multiplication	Division
Fractions	Computation	Cell 1	Cell 5
	Story Problems	Cell 2	Cell 6
Decimals	Computation	Cell 3	Cell 7
	Story Problems	Cell 4	Cell 8

Shaded areas indicate instructional coverage.

A comparison of the examples in Figures 1 and 2 suggests that a test may appear to have lower curricular validity when the comparison matrix contains a greater number of cells (e.g., 50% with four cells compared to 25% with eight cells). Judgment is required in determining of the number of cells to include in a content matrix for measuring instructional/test overlap. Dividing content into a large number of cells may preclude novel test items, because they will not match (overlap with) cells in which instruction has occurred. Additionally, with a matrix of very narrowly defined content in each cell, each cell may have only one possible item, precluding the development of alternate forms of a test covering all cells in the matrix. At the other extreme, if too few content matrix cells are used, only weak evidence of curricular validity can be obtained, at best. The optimal number of matrix cells will allow reasonable specificity of content while providing a sufficient range of possible test items for each cell in the content matrix. The goal is to retain the ability to test for application and understanding while still providing sufficient evidence to convince a court that the test is fair.

Once a content matrix of suitable size has been specified, the period of instructional time to be included in the quantification of overlap must be determined. The length of the instructional period to be included is important because most academic skills are taught over more than one year of instruction and are more readily mastered after repeated exposure. All else being equal, the more years of instruction that are included, the more likely the test will overlap with the instruction, and hence the greater the evidence of curricular validity.

Conjunctive vs. Compensatory Decisions

The test score is one piece—but not the only piece—of data on which a high-stakes decision may be made. To receive a high school diploma, a student also must obtain the required number of total credits, take the required number of courses in each subject area, and receive passing grades in all courses. Failure to satisfy any of these requirements will have the same result as failure to pass the diploma test—denial of a diploma.

The decision-maker can choose to combine the testing and other data required for obtaining a diploma by using either a conjunctive or a compensatory model. In the conjunctive model, each requirement must be satisfied in its entirety; outstanding performance in one area *cannot* compensate for poor performance in another. For example, a high test score cannot compensate for a failing grade in English, and an "A" in English Composition cannot compensate for a writing test score that is below the passing standard.

Effectively, in the conjunctive model, each requirement has its own passing standard and failure to meet any one of those standards results in failure to obtain the diploma. This model is the most common one used in diploma testing. To earn a diploma, students must meet all course requirements, achieve a minimum grade point average, complete the required total number of credits, *and* pass the graduation test. The conjunctive model also is used when a state sets separate passing scores for the reading, mathematics, and writing portions of the graduation test and requires the student to pass all three tests to receive a diploma.

Another way to combine data is to use a compensatory model. In the compensatory model, low performance in one area can be offset by outstanding performance in another. In the graduation context, a compensatory model would allow a student who scores just below the passing standard on the test to earn a diploma if the student's high school grade point average is above a specified level—e.g., 3.0. In another example of a compensatory model, the state establishes a total score passing standard for the graduation test, which combines performance on reading, mathematics, and writing subtests. Under this system, students with inadequate mathematics skills still may receive a diploma if their reading or writing scores are high enough to offset the low mathematics test scores.

Still another application of the compensatory model can occur within a single test. For example, if a reading test includes subtests measuring the ability to comprehend narrative and informational texts, the narrative and informational subtest scores may be combined to form a total score. If the passing standard is set on the total score scale, students who perform poorly on the informational subtest but do well on the narrative subtest may still pass the reading test.

To justify the use of a compensatory model, the decision-maker must be able to argue convincingly that achievement exceeding the passing standard in one subject can counteract an achievement deficit in another subject. For instance, the decision-maker would have to argue that students with superior math achievement should receive diplomas even if their reading ability is marginal. The counter-argument is that high school graduates must be competent in *both* mathematics and reading, because mathematics cannot be used to read a tax form and reading cannot be used to balance a checkbook.

One potential advantage of the conjunctive model is that the data may be collected sequentially, allowing decision-makers to stop collecting data as soon as the student fails to meet one of the requirements, because that substandard performance eliminates the person from further consideration. After all, there is no point in incurring the cost of collecting additional data that will not alter the decision. The compensatory model does not offer this advantage. All data must be collected on all persons, because a student can compensate for substandard performance on one measure by exceeding the standard on another measure. However, in testing applications with separate passing standards for each subtest (i.e., using the conjunctive model), the entire test (e.g., reading, mathematics, and writing subtests) may be given at one time to identify those areas requiring remediation and retesting—that is, the conjunctive model does not necessarily require a state to administer tests separately.

Technical Issues with Compensatory Models

If after careful consideration a decision-maker chooses to adopt a compensatory model, either within or across subtests, written documentation of the rationale for that decision should be prepared. The rationale should clearly describe the numeric scale to be used for the combined score, the weights to be assigned to each component of the combined score, and any specific rules governing the manner in which the scores are to be combined. Such documentation provides both clear communication to test takers/users and evidence of thoughtful deliberation in the event of a challenge.

For example, if a writing test contains 20 multiple-choice editing items and two essays each scored on a five-point scale by two raters, a decision-maker might create a combined writing score that is the sum of the multiple-choice raw score plus the four essay ratings. If students were required to achieve a total score of 24 or higher out of 40 total points to receive a diploma, compensation would be possible. That is, a student who correctly answered only half of the 20 multiple-choice items could still receive a diploma if the four essay ratings were all 4s [10 + 4(4) = 26]. Similarly, a student with all 2s on the essays could still receive a diploma with a multiple-choice score of 16 [16 + 2(4) = 24].

However, even when a decision-maker has carefully documented the rationale for the scale designed in connection with the adoption of a compensatory model, two major problems will remain: illusory weighting and lack of comparability for subsequent forms of the test. The lack of comparability problem occurs when a harder form of the test is given at a subsequent testing, making it more difficult for students to achieve the 24 points necessary for a diploma. This issue can be addressed with a technical procedure called equating. Equating is discussed in greater detail in a later section of this chapter.

To understand the illusory weighting issue, consider the previous example of a combination editing and composition writing test. Of the 40 total writing score points, 20 came from multiple-choice items and 20 from essay ratings. Thus, it appeared that the two sections of the writing test were equally weighted in the total score.

But suppose that in the first year in which the test is administered the statewide means and standard deviations of the multiple-choice and essay sections, respectively, were (10, 2) and (10, 4). Suppose further that two students who took the test were each good at either editing or composition, but not both, and scored two standard deviations above the mean in their strong areas and two standard deviations below the mean in their weak areas. Their respective scores would be as follows:

	Multiple-choice	Essays	Total	Result
Student #1	14	2	16	Fail
Student #2	6	18	24	Pass

Even though the multiple-choice section appears to count equally in the total score, the student who is stronger in composition passes, while the student who is stronger in editing fails. Student #2 is disadvantaged little by the below average performance in editing but helped greatly by the above average score on the essays. Similarly, Student #1 is significantly disadvantaged by the below average performance on the essays and helped little by the above average performance on the editing section. Thus, although the equal weighting of the editing and essay portions would appear to have created an advantage for neither student, an advantage actually accrued to the student who was good at composition. This result occurred because the standard deviations of the two sections of the writing test were unequal. The essay section had the larger standard deviation and therefore the greater weight in determining the final outcome.

Only if the standard deviations were equal would the weighting of the two sections be equal, as intended. For example, suppose the standard deviations of the two sections had both been 4. The scores of the two students would have been as follows:

	Multiple-choice	Essays	Total	Result
Student #1	18	2	20	Fail
Student #2	2	18	20	Fail

In this case, both students would receive the same total score and the same result. In summary, both the illusory weighting issue and the equating issue to be discussed subsequently suggest (1) the need for caution in adopting a compensatory model for high-stakes decisions and (2) the need to base all score reporting on standard score rather than raw score scales. However, the alternative conjunctive model for decision-making has the disadvantage of placing a heavy measurement burden on each individual measure. Individual measures tend to be less reliable, resulting in more frequent false positive and false negative decisions.

Setting Passing Standards

On the continuum of knowledge/skills from competent to incompetent, the passing standard has been judged to represent the minimal knowledge/skills judged important for recipients of a diploma or license. Experts may use several methods to make this judgment. The commonly used methods of setting passing standards employ different assumptions and methodologies and do not produce identical results. Consensus standards based on a combination of beliefs about what minimally competent candidates should be able to do and data that indicate what typical examinees are able to do are preferred. Although it is common to ask content experts to recommend a passing standard, such standards also must be adopted by the individual (e.g., state superintendent or commissioner of education) or group (e.g., state board of education) with the authority to choose a passing standard.

When setting a passing standard for a high-stakes test, states must balance two types of errors: false positives and false negatives. False positives are persons whose test scores equal or exceed the passing standard but who do not really possess the minimum knowledge/skills being tested. False negatives are persons who possess the minimum knowledge/skills tested but whose test scores fall below the passing standard. These errors are inversely related—decreasing one increases the other, and vice versa.

False positives and false negatives occur because tests are imperfect measures. Due to idiosyncracies in the interaction of individuals and tests, the test score represents the individual's true achievement plus or minus some error of measurement. If an individual were tested repeatedly with equivalent test forms, the resulting test scores, although different, would tend to cluster around the individual's true achievement.

The standard error of measurement is an estimate of the accuracy of measurement of a given test. The more consistently a test measures, the smaller will be the standard error of measurement. The standard error of measurement provides one indication of the amount of positive or negative change in an individual's score that might be expected if the individual were retested. Theoretically, errors for individuals are random; sometimes a person has a positive error resulting in a test score above the person's true achievement and sometimes a person has a negative error resulting in a test score below the person's true achievement. The random nature of measurement errors means that one cannot predict ahead of time which kind of error a person will have and that on repeated testings the error of measurement will not be consistently positive or negative, but will vary.

When examinees are allowed to retake the test several times to achieve a passing score, there is virtually no chance that a person with true achievement *above* the judged minimum passing standard will not pass after three attempts. However, 87.5% of those with true achievement *equal to* the judged minimal passing standard would pass after three attempts, and those with true achievement just below the passing standard would have a substantial likelihood of being judged competent after multiple attempts. Thus, with multiple retakes, the potential is much greater for false positives than for false negatives.

Some commentators argue that individual rights are most important and that no competent person should be denied a credential. To avoid the possibility that a marginally qualified candidate would fail to achieve a passing score on the first attempt, they argue that the actual passing score used to make decisions ("operational passing standard") should be substantially lower than the judged minimal standard determined by the standard-setting process. But the price of setting an operational passing standard far enough below the judged minimal standard so that all competent persons will pass is to risk the award of credentials to a significant number of incompetent persons. In such circumstances, the government interest in protecting the public has been sacrificed to entitle competent individuals to credentials. Some commentators argue that multiple retakes are enough protection against individual errors of measurement and that society should be more worried about credentialing incompetents than about borderline competent persons who must retake the test to pass.

Some commentators who have argued for a more flexible standard setting process have proposed awarding credentials to persons whose test performances fall within three standard errors of measurement below the level established in the standard setting study. This view is based on the reality that even the most widely used standard setting methodologies are not scientific, but judgmental, and thus are inherently inexact. It also is sometimes argued that this approach would legitimately ameliorate the well-documented adverse impact of many tests on members of historically disadvantaged groups.

The critic's rejoinder to this argument is that representatives from historically disadvantaged groups participate in the standard setting process and that knowledge of measurement error and student capabilities are already taken into account when passing standards are set at 50%, 60%, or 70% of a set of items that are believed to be important and attainable for all students. The critics would argue that any further lowering of the passing standard would dilute the required achievement to an unacceptably low level.

Suppose decision-makers believed that all students with high school diplomas should be able to add, subtract, multiply, and divide whole numbers. (For the moment, ignore the fact that other skills also might be tested.) Suppose further that a 50-item multiple-choice test were designed that included equal numbers of content-validated computation problems for each of the four operations. Finally, suppose a standard setting panel, with representation from historically disadvantaged groups and access to student performance data, set a consensus passing standard of 60% on this test, which had a standard error of measurement of 5 points. The raw score passing standard in this context would be 30 items correct.

According to the view that would further compensate for adverse impact by lowering the passing standard by three standard errors of measurement, the actual cut score would be 15. But on a four-choice multiple-choice test, a student with little knowledge who guessed at all of the items could be expected to get approximately 12-13 items correct. Critics of this procedure would argue that 15 is too close to a chance score to indicate meaningfully even minimal mastery of the skills tested and that a score of 30% correct (15 items out of 50 total) could be achieved on multiple retests by some students who really knew how to solve less than 30% of the problems. They also might argue that it is difficult to explain adverse impact on math computation items as a function of cultural bias.

The bottom line issue that decision-makers must confront in this debate is whether documented adverse impact should be addressed by lowering the passing standard to a level that will award credentials to a substantial number of individuals who clearly have not achieved the knowledge/skills being tested. Put another way, decision-makers must determine whether the information obtained from a test with such a low passing standard will improve the current system of awarding diplomas sufficiently to justify the substantial costs of administering such a testing program.

Equating

In a high-stakes testing program, the same test form cannot be used year after year. (See later sections on The Lake Wobegon Effect and Test Security.) Each administration of a high-stakes test requires a different form with a high proportion of new items. However, to be fair to students who take different forms of a test, states must equate each new form to a common score scale. By doing so, the state can ensure that the passing standard will remain constant for each new group of students tested or for those students who are retested after remediation. Thus, if a new test form is more difficult than the previous one, the number of items that the student must answer correctly to pass decreases. Conversely, if the new test form is easier than the previous one, the number of items the student must answer correctly to pass will increase. Another way to think of this is that when measuring student achievement, correct answers to hard items are worth more than for easy items.

In many testing applications, test forms are equated using a common item design, which means that the two forms of a test have a core set of identical items. But this method requires that the common items are unknown to the students who take each test form. Thus, it is vital to fair equating that the students taking the test the second year have not seen the common items from the previous year's test. Therefore, test forms and items cannot be released to educators for instructional purposes after testing. If items are released or test security is compromised, those items cannot be reused. If all items on a test form are new, it is still possible to determine the equivalent passing standard, but the equating may be less accurate and the number of items that must be answered correctly to pass may change significantly from year to year.

Pre-equating

When equated raw score passing standards change significantly from one year to the next, those who do not understand equating may question the validity of the test scores. One way to avoid significant changes in the raw score passing standard from form to form is to pre-equate each form before it is administered. Pre-equating uses pretest data to estimate the relationship between scores on a new test form and scores from previously administered test forms. Pre-equating provides an estimate of the raw score passing standard for any set of pretested items with item statistics adjusted to a common scale.

If the pre-equating suggests that the raw score passing standard for a particular form will be significantly different from that of previous forms (e.g., five points higher or five points lower), some appropriately harder or easier items (which meet all other selection criteria including match to the objectives or test specifications) can be substituted until the raw score passing standard is close to that for prior forms.

Pre-equating avoids the public relations problem of explaining to lay persons, who are unable to detect the differences in difficulty between two test forms by inspection, that the state has not changed the passing standard. When fewer correct answers are required to pass the test, the public may believe that the state has lowered the standard, even though it has not.

Flawed pretest data can adversely affect pre-equating estimates. Concerns about cost and item security may lead test developers to pretest items on small samples of students. Smaller samples produce less accurate estimates of item statistics than larger samples. If pretest items are administered separately rather than within operational forms, students may not try hard, because they know that the separate pretest items do not count. When students are not highly motivated, item statistics are distorted and some difficult pretest items may actually be relatively easy for students when they are administered later as part of a high-stakes test form. However, even when not accurate enough for pre-equating, pretest data still can be used to detect flawed items.

National Comparisons

States that adopt diploma testing requirements often have dual purposes for testing. Although they want to ensure that each student who receives a diploma has achieved specific skills, they also may want to aggregate the data for evaluation purposes. In particular, a state or district may want to compare the achievement of its students with the achievement of students nationally. Favorable comparisons and/or the demonstration of improvement over time can bolster claims of outstanding educational systems and attract new businesses to the state. However, if not done carefully, such comparisons can be misleading at best.

Some states have tried to scale the state-developed test to a nationally normed, standardized achievement test. Scaling is a measurement technique that links performance on two different tests by determining corresponding scores on the two tests that represent equivalent levels of performance. Scaling is most accurate when the two tests are parallel forms (developed from the same test blueprint or set of objectives) and of approximately equal difficulty (means and standard deviations about the same). The most common designs for such scaling are to administer both tests to the same group of students in counterbalanced order or to administer the two tests to two randomly equivalent groups of students.

Once scores on the statewide test have been converted to their equivalents on the national achievement test, national percentile ranks can be assigned to each score point on the statewide test. Then, for students who have been administered the statewide test but not the national achievement test, one can estimate what their national percentiles would have been if they had taken the national test.

A major problem with this method of obtaining national norms is that the statewide and national tests typically differ significantly in content and difficulty. While the statewide test focuses on state-specific instructional objectives, the national test focuses on a much broader content domain common to textbooks and curricula from multiple states and grade levels. Thus, the coverage of the national test is typically much broader than that of the statewide test, and the national test contains fewer items per objective. In addition, the national test usually has a broader range of item difficulties and a greater number of more difficult items than the statewide test. Together, these differences decrease the accuracy of the scaling, because the statewide and national tests are not parallel forms and are testing somewhat different content at different levels of difficulty.

When states began scaling their state-developed tests to nationally normed standardized achievement tests, the conversion tables obtained from the scaling design in the first year of the program were used for several years thereafter. This practice allowed districts to appear to improve significantly nationally when only minor changes in performance had actually occurred. By concentrating instruction on a few specific skills, educators could significantly increase scores on the statewide test. As scores increased on the statewide test, so did the corresponding national percentile ranks. However, the increased achievement represented by instruction on the statewide objectives covered only a small fraction of the skills tested on the national test. Thus, if those same students had taken the national test, their performance would have changed only slightly. The actual national percentile rank corresponding to performance on the national achievement test therefore would have been much lower than that estimated from the statewide test performance.

Using the procedures described above, the states actually were trying to obtain national norms for their statewide tests. The better way to obtain such information would be to give the statewide test to a representative national sample of students. However, such a massive data collection effort would be prohibitively expensive for an individual state. As a result, most states have opted for dual testing to meet their twin goals of specific standards for high school graduates and comparisons to national performance. Because of this practice, tests for awarding diplomas have returned to a focus on those skills deemed necessary for all high school graduates.

The Lake Wobegon Effect

Related to the national norms issue is a phenomenon known as the Lake Wobegon Effect. The Lake Wobegon Effect was first described by Dr. John Jacob Cannell, then a physician from West Virginia. The Lake Wobegon Effect derives its name from Garrison Keillor's mythical town in Minnesota "where all the women are strong, all the men are good-looking, and all the children are above average."

By definition, average means "middle" or "midpoint" and requires that some schools be above the average, while others are below the average. But in Dr. Cannell's surveys of the 50 states and large urban districts that were administering nationally standardized achievement tests, *all* reported themselves as being above average. Aside from the impossibility that all states and districts were above average, compelling evidence suggested that significant numbers of students lacked basic skills and were illiterate. Thus, Dr. Cannell charged that the rosy picture that states were painting of their above-average national performance was at best misleading and at worst fraudulent.

The Lake Wobegon Effect had several causes. A major underlying cause was the accountability movement, which provided monetary rewards to schools and teachers whose students performed well on standardized tests. The pressure to do well led teachers to spend weeks before the test drilling students on the specific objectives tested. Meanwhile, these teachers neglected the remainder of the curriculum, from which the specific objectives tested had been sampled.

In some cases, administrators and teachers went a step farther and inappropriately taught the actual test items or prepared drill worksheets closely paralleling the test questions. Such tactics were effective in increasing students' scores because (1) the original norms were based on samples of students who had not been specifically taught the test, and (2) many schools reused the same forms of the standardized test year after year. Although the standardized test was supposed to indicate how well students had learned a broad, national curriculum by systematically sampling content from that broad domain, the test scores actually became measures of educators' efforts to teach (appropriately or inappropriately) only the specific content included in the small sample of test questions. Thus, it was not surprising that performance could exceed the average of a national group of students who had been instructed on the broad curriculum and who had not received any special preparation for the test.

Some educators believed that such inappropriate test preparation was justified because of the punitive character of the accountability movement. Such thinking encouraged expanded use of inappropriate test preparation activities and creative juggling of statistics to produce the desired effect. The perception that state accountability actions were punitive also caused ripple effects at the local level (e.g., tying teacher evaluations to student performance) that encouraged further inappropriate activities.

As states became more aware of the potential for inappropriate test preparation activities, they began to scrutinize the testing enterprise much more closely. Other abuses were quickly identified. These abuses included helping students during the test, changing answers on students' answer sheets, practicing with the actual test items or parallel forms of the test just prior to test administration, encouraging special education and other low-achieving students to stay home on the day of testing, copying or taking test booklets, obtaining information about the test questions from outside sources, allowing students extra time to finish the test, and sharing test content with students taking the test later during makeup periods.

Test Security

As a counter-measure to the inappropriate test preparation activities associated with the Lake Wobegon effect, states have greatly increased test security for both nationally standardized tests and for their statewide tests. The goal is to make the testing as fair as possible for everyone so that no student receives an unfair advantage. No security procedures are one hundred percent effective, but states that have implemented such procedures report that testing irregularities have been greatly reduced.

Unfortunately, test security procedures are costly. The costs include added expenses for the following deterrents: numbering and shrinkwrapping test booklets; sending materials by special carriers just a few days before testing and picking them up again immediately afterward; developing special forms for designated district officials to sign for receipt, handling, storage, and return of testing materials; randomly auditing schools to check on compliance with mandated test administration procedures; checking answer sheets for excessive erasures; scrutinizing gains that appear to be "too good to be true"; investigating reports of inappropriate test preparation activities and sanctioning educators who engage in them; changing significant numbers of test questions from year to year; and holding educational seminars that provide notice of inappropriate test preparation activities and suggestions for appropriate test preparation. Only the larger testing programs are able to afford all of these measures; smaller programs must rely on educational efforts and ethical deterrents. States must be vigilant, because if inappropriate practices become widespread, the credibility of the testing program suffers and the legal requirement of fundamental fairness to all students is violated.

In addition to implementing procedures to deter inappropriate testing activities, states face special problems in maintaining secure test forms. To avoid special advantages due to educators' remembering the content of items on a given test, a substantial number of new items are needed for each year's test. New items are costly to develop and pretest, and additional resources are needed to typeset, print, distribute, and retrieve new test forms each year.

Sampling Objectives

Some critics of diploma tests have charged that such tests narrow the curriculum because teachers under pressure to do well teach only the specific skills tested and neglect the remainder of the curriculum from which the test objectives were sampled. To address this concern, some states have considered sampling objectives on a rotating basis from year to year.

For example, if a mathematics curriculum contained 100 objectives and 25 objectives could be sampled on a single test form, 25 objectives could be tested in the first year, 25 in the second year, 25 in the third year, and the remaining 25 in the fourth year. The theory is that if teachers do not know which 25 objectives will be tested each year, they will be more likely to teach all 100 objectives each year. As a sample of the total curriculum, each year's test would then provide a more accurate picture of students' mastery of the total curriculum. For equating purposes, slight modifications in this design can be made so that adjacent years' tests have the common items needed to place all test forms on a common scale.

Instead of dividing the 100 objectives into four mutually exclusive and exhaustive sets of 25, another approach is to select a stratified random set of objectives for each test form. To implement this approach, each objective must be classified according to its content and weighted according to its importance. Twenty-five objectives are then randomly selected based on both content and relative importance. Each objective has a chance of being represented on each test form, but the objectives do not all have an equal chance of being represented.

Equal Opportunity vs. Equal Outcome

Considerable debate has focused on whether the guarantee of equal opportunity in education implies equal outcomes. In striking down segregation by race or disability, the U.S. Constitution and federal laws have guaranteed all students equal access to education but have not guaranteed any specified level of achievement.

In his dissent to the appeals court's denial of a rehearing in the initial *Debra P.* case, Judge Tjoflat characterized a high school diploma as a reward to be earned, not an entitlement of educational seat time. Judge Tjoflat further suggested that constitutional due process requirements should be applied to access to education and not to academic standards required for graduation. Judge Hill, who also dissented, stated that it was unreasonable for a student who had not learned minimal skills to expect the courts to mandate that a diploma be awarded anyway.

Judges Tjoflat and Hill further argued that granting diplomas to students who had not learned the specified content and skills sanctioned unequal treatment, because it rendered the diploma a worthless piece of paper and perpetuated stereotypes of African-American underachievement. They suggested that real equality could be achieved only by maintaining educational standards, identifying students who have not learned the specified content and skills, and providing remedial instruction. Indeed, by the time the ***Debra P.*** case concluded, and after receiving remedial instruction, over 90% of African-American graduating seniors had passed the graduation test.

The Lowest Common Denominator

Not everyone is equally talented academically. Given a fixed amount of time, some individuals will achieve more than others. As a result, some critics have charged that diploma testing has reduced what is taught in the public schools to the lowest common denominator of what the least prepared students are able to learn.

Other educators argue that without meaningful standards for judging achievement, some students might just exist in school without getting the attention and help that they so desperately need. These educators suggest that a wiser use of the message from high-stakes tests might serve the public interest better than killing the messenger and pretending that the problem does not exist. These educators argue further that the interpretation of test scores must take into account the large portions of a student's life over which schools and training programs have no control. When test scores are interpreted reasonably, they suggest, states can more effectively and fairly allocate resources and ensure student achievement of important skills.

To put this issue in perspective, one must remember that it is not the *test* that is high-stakes, but the *decisions* that are made based on the test. These decisions will be made whether or not the test is given. In general, the more data a decision-maker uses to make a decision, the better the decision. Assuming that the decision-maker wants to award diplomas only to students who have learned specified content and skills, and assuming that the test is valid, the decision-maker will make better decisions with the test information than without it. Even though a test might be improved given additional resources, it is important to remember that the informal, subjective judgments that might be used in place of a test may result in substantially more erroneous decisions, because human judgments tend to be even more inaccurate and biased than test scores.

Differentiated Diplomas

For educators who believe that a high school graduate should have more than minimal skills, a new concept in diploma testing has been proposed. The proposed new plan calls for differentiated diplomas depending on passing different tests or passing the graduation tests at different passing standards. The idea is to retain the minimal standards for awarding diplomas but to give higher-achieving students the opportunity to demonstrate superior achievement and to recognize that higher level of achievement with a special endorsement on the diploma.

Critics argue that such schemes are simply new ways to discriminate against historically disadvantaged groups, since few members of these groups may be able to achieve the special endorsements. Given the likelihood of disparate impact and the tendency for historically disadvantaged groups to be overrepresented in special education programs but underrepresented in gifted programs, many states have been deterred from serious consideration of differentiated diplomas. However, a few states and districts are considering endorsed diplomas for students who pass state or district achievement tests in specified academic subjects.

Some educators prefer standardized diplomas with endorsements recognizing exceptional achievement rather than different diplomas or the withholding of a diploma in cases of substandard performance. However, if courts view endorsements as denied property rights, challenges similar to *Debra P.* may occur. Viewing the endorsement as a relevant property right may be reasonable if employees condition jobs on endorsements rather than possession of a diploma. Nonetheless, endorsements may be more viable politically, since students who fail the endorsement tests or do not take the tests because they are in special education programs still can receive unendorsed high school diplomas if they satisfy all course or IEP (Individualized Educational Program for students with disabilities) requirements.

Naming the Test

A relatively minor point, but an important public relations issue, concerns the name given to tests used to award diplomas. The Florida Functional Literacy Examination was criticized in part because students who failed were cast as "functional illiterates." A core issue in the *Debra P.* lawsuit was the challengers' assertion that the state has an obligation to demonstrate the *predictive validity* of a test that purports to measure "functional literacy." This nomenclature was so inflammatory and caused such heated debate that the state renamed the test the State Student Assessment Test Part II (SSAT II).

Any test name that suggests that it measures a psychological construct such as "functional literacy" will precipitate debate about the definition of that construct and appropriate ways to measure it. Although Florida's intent was to make the test relevant to tasks encountered in everyday life, it instead attracted disagreement about which skills are essential for everyday life and whether life skills were part of the curriculum in Florida schools. To avoid possible legal wrangling over terminology and pseudo-issues, the wiser course may be to give the test a neutral, general title and allow the skills that it measures to be defined by the test specifications and curricular frameworks or lists of specific objectives.

Summary

Most states have or are considering high-stakes statewide testing programs. Although this activity had been decreasing in the last few years, the performance assessment movement appears to have generated renewed interest in high-stakes statewide testing. Public support for accountability is significant, and a variety of special interest groups are closely scrutinizing public education. When high-stakes statewide testing programs include tests for awarding diplomas, handling potential and actual legal challenges requires careful planning, knowledge of legal and professional standards, comprehensive documentation, and a decision-making process that is procedurally fair.

Public relations programs that keep constituent groups informed and involved in the process at all stages may do much to foster workable compromises and forestall formal court challenges. But when educators, legislators, policymakers, and representatives of special interest groups are unable to reach consensus, aggrieved groups often will not hesitate to seek judicial remedies. Awareness of potential challenges and the options available to states should assist them in constructing more defensible testing programs.

Recommendations for Developing and Implementing Legally Defensible Statewide Tests for Awarding Diplomas

The following recommendations pertain generally to traditional diploma testing programs for general education students. Issues related to the newer performance assessments, item/test bias reviews, and testing accommodations for disabled persons are discussed in subsequent chapters. These recommendations are drawn from a variety of sources, including the information presented in this chapter, professional standards for testing, the experiences and professional judgment of the author, and common sense.

(1) Establish a technical advisory committee to advise the state agency (e.g., department of education) and state board on all policy matters and decisions related to the high-stakes assessment program.

(2) Codify all major policies in administrative rules formally adopted by the state board. At minimum, the state board should officially adopt curricular frameworks, test forms, accommodations policies, test security policies, and passing standards.

(3) If not already provided for in a state tort claims act, consider sponsoring legislation to provide limited immunity to professionals in the state who assist the state agency in the development of the testing program.

(4) Involve representatives of major constituencies (e.g., teachers, unions, administrators, disabled persons, historically disadvantaged groups, business, and parents) in advisory groups providing input on testing policies and content.

(5) Provide districts and students with two to four years' advance notice of the content and format of the assessment program. Lists of specific curricular objectives, sample questions, and suggestions for appropriate test preparation provided by the state agency are helpful. Regional meetings to disseminate information and solicit input are also desirable.

(6) Provide at least as much notice to special education students and other special populations about the policies regarding testing that will apply to them as is provided to general education students.

(7) Develop and follow a written testing accommodations policy sufficiently in advance of the first testing date.

(8) Provide multiple opportunities for passing the test and ensure that remediation is available to those who do not pass.

(9) Document that the content being tested is being taught by the school districts in the state (curricular validity) sufficiently in advance of the date when diplomas will first be denied based on the tests, so that students have an adequate opportunity to learn the tested material. Trial administration of test forms one or more years prior to the implementation of the diploma sanction can help satisfy both the notice and curricular validity requirements.

(10) Establish passing scores as consensus standards based on a combination of content judgments and performance data.

(11) Provide a phase-in period for new curriculum before including it on the test.

(12) Provide written materials and workshops for assisting districts in interpreting and using test score information.

(13) Design score reports that communicate effectively to those with minimal knowledge of assessment.

(14) Implement the following test security guidelines:

    (A) Ship test booklets so that they arrive only a few days before testing. Require a responsible administrator to sign a form acknowledging receipt and assuring that the materials will remain locked in a storage area with very limited access.

    (B) Allow only the minimum necessary time for testing and require all sites to test on the same day(s).

    (C) Require all testing materials to be returned immediately after testing.

(D) Seal and number all test booklets and shrink wrap bundles of test booklets.

(E) Require written assurance from test administrators at each site that test booklets were opened only by examinees when told to do so during testing and that no booklets were photocopied.

(F) Require test administrators to account for all testing materials before examinees are allowed to leave the room for lunch breaks or at the conclusion of testing.

(G) Arrange for multiple proctors in each testing room and allow only one student at a time to leave during testing.

(H) Have all test administrators keep records of irregularities at the test site.

(I) Investigate all reports of breaches of test security and sanction those involved in confirmed incidents.

(J) Randomly audit test sites unannounced to ensure that proper procedures are being followed.

(K) Request that the legislature enact a statute or the state board adopt an administrative rule defining inappropriate test preparation activities, providing sanctions for individual educators who engage in inappropriate test preparation activities or cheating, and giving the state agency authority to investigate and impose sanctions.

(L) Examine answer documents for tampering, excessive erasures, copying, and other signs of cheating. Screen group statistics and repeat testers for unusually large performance gains. Use suspicious findings to trigger appropriate investigations.

(M) Where identity may be an issue, require each examinee to produce photo identification, sign the answer document at the beginning of each testing session, or place a thumb print on the answer document. However, these procedures may significantly increase administration time and expense.

(15) Seek technical assistance early in the testing program to design data collection for equating that will ensure that the achievement required to attain the passing standard remains constant from year to year.

(16) Follow professional standards in all technical matters, including, but not limited to, item development, item selection, validity, reliability, item bias review, equating, scaling, setting passing standards, test security, accommodations, test administration, scoring, and score reporting.

(17) Carefully consider the advantages and disadvantages of setting separate passing standards for each content area tested (e.g., reading, mathematics, and writing) or setting a single passing standard based on a composite total score. Involve relevant constituencies in the standard-setting process.

(18) Designate a state agency spokesperson to make all official announcements and comments about the testing program. Caution all state employees not to make unsubstantiated statements regarding what the test measures or inferences that can be made from test scores.

(19) Provide thorough training for members of item writing, standard setting, content review, bias review, and scoring committees.

(20) Consult with the attorney general's office or independent counsel regarding statutory requirements and potential litigation. Detailed documentation of all actions and policies should be available for review. Such information also may be accessible to the public through freedom of information requests. Exemptions may need to be sought for secure test materials.

(21) Designate trained state agency personnel to provide continuous and comprehensive supervision and interaction with all contractors for the testing program.

(22) Choose a neutral name for the test that does not include any constructs for which there could be debate and strong disagreement about their meaning. For example, The (state name) Graduation Test.

References

Cases and Statutes

Anderson v. Banks, 540 F. Supp. 472 (S.D. Fla. 1981), reh'g 540 F. Supp. 761 (S.D. Ga. 1982).

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), aff'd in part, rev'd in part, 644 F.2d 397 (5th Cir. 1981); on remand, 564 F. Supp. 177 (M.D. Fla. 1983), aff'd, 730 F.2d 1405 (11th Cir. 1984).

U.S. Constitution, Amendment XIV, section 1.

Articles and Other Resources

American Psychological Association (1985). *AERA/APA/NCME Standards for Educational and Psychological Testing.*

Calfee (1983). Standards, Evidence and Equity: Implications of the 1983 Debra P. Decision. *Educational Measurement: Issues and Practice, 2*(4), 11.

Cannell (1988). Nationally Normed Elementary Achievement Testing in America's Public Schools: How All 50 States Are Above the National Average. *Educational Measurement: Issues and Practice, 7*(2), 5.

Ebel & Frisbie (1991). *Essentials of Educational Measurement*, 5th ed.

Fisher (1983). Implementing an Instructional Validity Study of the Florida High School Graduation Test. *Educational Measurement: Issues and Practice, 2*(4), 8.

Gaddy (1988). High School Order and Academic Achievement. *American Journal of Education, 96*, 496.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education.* Washington, DC: Author.

Lee & Rong (1988). The Educational and Economic Achievement of Asian-Americans. *The Elementary School Journal, 88*(5), 545.

Linn (1983). Curricular Validity: Convincing the Courts That It Was Taught Without Precluding the Possibility of Measuring It. In G. Madaus (Ed.), *The Courts, Validity, and Minimum Competency Testing*, 115.

Lerner (1986). Student Self-Esteem and Academic Excellence, *The Education Digest, 52*, 32.

Mehrens (1986). Measurement Specialists: Motive to Achieve or Motive to Avoid Failure? *Educational Measurement: Issues and Practice*, *5*(4), 5.

Mehrens & Kaminski (1989). Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?, *Educational Measurement: Issues and Practice*, *8*(1), 14.

Mehrens & Lehmann (1991). *Measurement and Evaluation in Education and Psychology* (4th ed.).

Mehrens & Phillips (1986). Detecting Impacts of Curricular Differences in Achievement Test Data. *Journal of Educational Measurement*, *23*(3), 185.

Millman (August/September 1989). If at First You Don't Succeed: Setting Passing Scores When More Than One Attempt is Permitted. *Education Researcher*, *18*, 5.

Phillips (1991). Diploma Sanction Tests Revisited: New Problems from Old Solutions. *Journal of Law and Education*, *20*(2), 175.

Popham (1983). Task-Teaching Versus Test-Teaching. *Educational Measurement: Issues and Practice*, *2*(4), 10.

Shepard (1987). *A Case Study of the Texas Teacher Test: Technical Report*.

Thorndike, Cunningham, Thorndike, & Hagen (1991). *Measurement and Evaluation in Psychology and Education* (5th ed.).

Chapter 3

Potential Bias Against Historically Disadvantaged Groups

Overview

When members of a historically disadvantaged group (e.g., African-Americans, Hispanics, Native Americans, females) perform less well on a test item than majority group members of equal ability, the test item may be biased against members of the historically disadvantaged group. The *Golden Rule* procedure is a discredited remedy for addressing potential test item bias against historically disadvantaged groups. This procedure was part of an out-of-court settlement in a lawsuit challenging an Illinois insurance licensure examination. Although no court has ever mandated the *Golden Rule* procedure and measurement experts agree that it is an inappropriate method for identifying potentially biased test items, those who challenge testing requirements often pressure policymakers to adopt it.

This chapter traces the legal history of the *Golden Rule* settlement, explains why measurement experts have discredited the *Golden Rule* procedure, discusses appropriate alternative procedures for detecting and remedying test item bias, and examines attempts to extend the *Golden Rule* procedure in other legal contexts. This chapter also describes the legal arguments commonly used to assert that a test discriminates against a historically protected group, discusses applicable legal and measurement professional standards, and provides an example of the potential detrimental effects on test validity of a *Golden Rule*-type procedure.

Terms

adverse impact
differential item performance
disparate treatment
disparate impact
fundamental fairness
intent to discriminate
job analysis
test specifications

Cases

Allen v. Alabama State Board of Education—challenge to teacher licensure test that sought to impose a more stringent version of the *Golden Rule* procedure.

Golden Rule Life Insurance Co. v. Mathias—challenge to Illinois insurance licensure test for which the settlement mandated a procedure for identifying and remedying potential item bias against historically disadvantaged groups.

Legal Issues

disparate treatment/disparate impact
EEOC *Uniform Guidelines on Employee Selection Procedures*
Fourteenth Amendment equal protection
Fourteenth Amendment procedural due process
Fourteenth Amendment substantive due process
Golden Rule remedy and extensions
job relatedness
Title VII of the Civil Rights Act of 1964

Measurement/Educational Issues

APA/AERA/NCME *Standards for Educational and Psychological Testing*
appropriate bias detection procedures
content validity
differential item performance
distortion of test specifications
potential test vs. item bias

Key Questions

(1) What is test item bias and why is it important?

(2) What is the *Golden Rule* procedure and what does it require?

(3) Why have measurement experts discredited the *Golden Rule* procedure?

(4) What are the characteristics of appropriate alternative procedures for detecting and eliminating potential test and item bias against historically disadvantaged groups?

(5) Why is the *Golden Rule* procedure *not* a legal requirement in any jurisdiction?

Test developers recently have become concerned about potential bias in test items, while members of historically disadvantaged groups who score poorly on standardized tests have alleged that their low test scores are due to such bias. Specifically, some members of historically disadvantaged groups believe that test items reflect a white, middle-class culture that discriminates against historically disadvantaged groups whose culture and life experiences differ from those of the dominant population.

Sometimes the potential bias seems obvious, such as when urban students are asked about farm animals that they have never seen. In other cases, the potential bias may be more subtle, such as when a vocabulary word or term is not common or has a different meaning within the historically disadvantaged group's culture. But even reviewers from historically disadvantaged groups sometimes are unable to explain precisely why one item appears biased against a particular group and another item does not. For example, similar percentages of minority and majority students may answer the question 28 + 63 = ? correctly, while a significantly greater percentage of majority than minority students correctly answers 26 + 45 = ?

Moreover, advocates for historically disadvantaged groups often allege that the differential performance between majority and historically disadvantaged groups is a function of the different and inferior education received by historically disadvantaged students. They believe that majority students have had a greater opportunity to apply the tested skills than historically disadvantaged students. Whether or not this hypothesis is correct, the question addressed in this chapter remains the same: Is the test developer precluded from measuring mastery of such skills. In other words, if the education or social environment of historically disadvantaged groups is discriminatory, does it follow that the test is biased? Or is the test merely the messenger of bad news about the inferior opportunities afforded historically disadvantaged groups?

These questions raise the issue of how one should define bias in test items and what policies should be developed for dealing with bias. If a test item is determined to be biased according to an agreed-upon definition, the policymaker has two choices: delete the item from the test, or address the underlying cause of the bias in the educational or social environment. The former remedy may be more easily accomplished in the short run, but it also may have long-term negative consequences for maintaining standards and fulfilling the state's duty to protect the public from incompetent practitioners. The latter solution requires a longer time frame and greater resources, and some historically disadvantaged group members may find themselves caught in the middle during the transition.

Recent efforts to eradicate bias in testing have emphasized screening test items for bias and eliminating the offending items. In most cases, these item bias procedures focus on differential performance between majority and historically disadvantaged groups. This approach makes sense if the groups being compared have similar abilities, because one might then hypothesize that the differential performance is caused by something in the item that cues majority group members or misleads historically disadvantaged group members.

However, some reformers have been drawn to simplistic procedures for protecting against bias that examine differences in performance without controlling for ability. Such data formed the basis for the settlement in *Golden Rule Life Insurance Co. v. Mathias*, in which those challenging the test sought to require the test developers to choose items answered correctly by African-Americans and Caucasians in approximately equal percentages. Although the *Golden Rule* remedy has never been mandated by any court and has been discredited by measurement professionals, reform groups continue to cite it when lobbying legislators, policymakers, and test developers to address potential bias in high-stakes tests. To provide a more complete understanding of this case and its significance in future litigation, the following sections describe the lawsuit, the legal challenges, the settlement, the measurement issues, and the aftermath.

The *Golden Rule* Lawsuit

The *Golden Rule* case involved an insurance licensure exam developed by the Educational Testing Service (ETS) and administered in the state of Illinois. African-American applicants challenged the licensure test, alleging that they had failed the test because it contained biased items. After years of procedural battling, the case was settled out of court.

The event that triggered concern about the fairness of the test was a steep drop in passing rates after the state adopted a new version of the licensure exam. In 1975, under contract with the Illinois Insurance Department, ETS developed a new insurance agent licensure exam based on specifications approved by the Insurance Department. When this new test was administered, the passing rate dropped from a prior range of 60-70% to approximately 30%. The next year, the insurance licensure test was revised and the passing rate rose to the 70-75% range.

But J. Patrick Rooney, Chief Executive Officer of the Golden Rule Life Insurance Company, believed that a disproportionate number of African-Americans were still failing the licensure exam. On behalf of several persons who had failed the test, the company filed suit against ETS and the Illinois Insurance Department to halt further testing. The complaint alleged that the individuals who had not passed the test were fully qualified insurance agents recruited by the company who had to be discharged solely because they had not passed the required licensure exam.

The company obtained a temporary injunction to halt testing. This injunction was based on a defect in the fee structure for the test, not on the substantive issue of bias. The defect was corrected, the test was revised again, and the court dismissed the lawsuit as no longer relevant. But the company refiled the lawsuit. This time, the complaint sought damages for prior discrimination under the old test in addition to an injunction to halt further administration of the new test.

The company alleged that both tests violated the Fourteenth Amendment to the U.S. Constitution and portions of the federal Civil Rights Act of 1964. The company cited the following factors as evidence of discrimination under federal law: (1) a larger percentage of African-Americans than Caucasians failed the licensure examination each time it was administered; (2) a larger percentage of African-Americans failed such examinations developed and administered by ETS in other states; (3) the percentage of Caucasian insurance agents and brokers in Illinois was higher than the percentage of all Caucasians in Illinois; and (4) the percentage of historically disadvantaged group members who passed the licensure test was lower than the percentage of all historically disadvantaged group members in Illinois. Statistically, the difference in passing rates for African-Americans and Caucasians—or the disparate impact of the test—ranged from 15% to 25% on various versions of the licensure test.

The court again dismissed the case, citing two major reasons: (1) ETS, as a private contractor, was not a state actor, which is required for constitutional and civil rights challenges, and (2) the company had failed to allege intentional discrimination, which is required for a violation of the Fourteenth Amendment. The company appealed the case, asking the appeals court to rule on three major issues: (1) whether ETS qualified as a state actor, (2) whether the complaint contained all of the elements required for a constitutional challenge, and (3) whether the company could sue under the federal Civil Rights Act.

The appeals court ruled in favor of the company on all three issues. Of greatest significance was its finding that ETS was a state actor. The court based this finding on a contract between ETS and the Illinois Insurance Department that afforded ETS a great deal of autonomy and authority in developing and administering the licensure tests. The court held that there was a very close nexus between ETS and the Department, that they were effectively partners in the testing endeavor, and that ETS could not claim to be a mere bystander. In addition, the court seemed to consider the company's argument that, because the company was willing to hire the recruits who had failed the licensure test, the licensure test developed and administered by ETS was is fact the sole factor preventing these recruits from being employed.

Note that all three issues addressed by the appeals court involved the procedural aspects of filing a lawsuit. Finding in the affirmative on these issues simply allowed the suit to proceed to trial; it neither supported nor refuted the substantive claim of discrimination. Indeed, the court expressed doubt that the claims of discrimination could be sustained at trial, but said that the company was entitled to its day in court. This outcome was justified, in part, because there was no dispute about the state's authority to protect the public by licensing insurance agents and because the neutral statute mandating the test indicated no intent to discriminate against any group. Thus, proof of intent to discriminate, required in a Fourteenth Amendment challenge, would depend on all of the surrounding facts and circumstances that could best be left to evaluation at trial.

Legal Issues

The licensure testing program in Illinois faced three potential legal challenges: violations of (1) Title VII of the Civil Rights Act of 1964 and the associated Equal Employment Opportunity Commission *Uniform Guidelines on Employee Selection Procedures* (*Uniform Guidelines*); (2) the equal protection clause of the Fourteenth Amendment; and (3) fundamental fairness under the substantive due process requirement of the Fourteenth Amendment. Title VII challenges are unique to employment testing applications and will be discussed in greater detail below. Equal protection and fundamental fairness under substantive due process were discussed in detail in Chapter 2 and will be reviewed only briefly here.

Title VII and the EEOC Uniform Guidelines

Title VII of the Civil Rights Act prohibits racial discrimination by employers. Title VII permits nondiscriminatory employment testing provided the test meets relevant professional standards. Because Title VII requirements are targeted at employers, it is questionable whether they would apply in a licensure context. However, because Title VII also prohibits discriminatory practices through contractual arrangements, it is arguable in this case that Title VII might apply to the quasi-contractual nature of the Illinois relationship. But because of the ambiguity regarding the application of Title VII to licensure tests, the *Golden Rule* case also challenged the insurance licensure test based on alleged violations of other sections of the Civil Rights Act.

In licensure testing, states require applicants to pass a test to obtain a professional license. The state derives its authority to impose licensure requirements from its power to protect the public welfare. The purpose of the licensure test is to protect the public from incompetent practitioners. Therefore, unlike an employment test, which is designed to distinguish levels of ability so that the most qualified persons can be selected, a licensure test merely certifies a minimal level of competence. Effectively, licensure applicants are classified into two groups: competent and not competent. All persons classified as competent (meaning that they have attained the minimal level of competence or a passing score on the licensure test) receive the same license to practice their profession.

When Title VII applies, litigants can follow two avenues to make a legal challenge: a "disparate treatment" or a "disparate impact" theory. Disparate treatment requires discrimination with the intent to disadvantage the particular individual who filed suit. Disparate impact requires only evidence of a statistically significant differential effect of the challenged practice on the historically disadvantaged group of which the complainant is a member.[1] Thus, in challenging a test, a disparate impact challenge under Title VII can proceed if there is evidence that significantly more African-Americans than Caucasians have failed the mandated test. Because few test users admit to discrimination or design instruments with the active intent to discriminate, discriminatory effect is a more common challenge than discriminatory intent.

Evidence of intent to discriminate is not required under a disparate impact challenge, because when Title VII was passed, Congress indicated that past societal discrimination was the intentional discrimination that the legislation was designed to remedy. Thus, individual complainants have been absolved of the responsibility of showing that the particular entity being sued intended to discriminate. The existence of differential group performance (or adverse impact) together with the assumption that the differential was caused by past discrimination by other entities is enough to establish the case. Once the complainant has demonstrated the adverse impact of a test, the burden shifts to the test user to validate the test.

The federal Equal Employment Opportunity Commission (EEOC) is responsible for enforcing Title VII requirements. Under its enforcement authority, the EEOC has developed the *Uniform Guidelines*, which delineate validation requirements applicable to employment test uses that have a disproportionately exclusionary impact on members of legally protected racial, ethnic, and gender subgroups ("adverse impact").

---

[1] Note that in most cases differences satisfying the four-fifths rule (discussed later) will be statistically significant.

The EEOC *Uniform Guidelines* require that adverse impact in testing be eliminated or, if there are no less discriminatory alternatives, that the test user demonstrate that the knowledge and skills being tested are job related. The terms "disparate impact" and "adverse impact" are synonymous. In determining whether a test has disparate or adverse impact, the *Uniform Guidelines* use the "four-fifths rule." This rule presumes adverse impact if the success rate for the historically disadvantaged group is less than 80% (or four-fifths) that of the majority group. For example, if 60% of Caucasian applicants obtain passing test scores and are hired, the test would be presumed to have adverse impact if fewer than 48% of African-American applicants obtain passing test scores and are employed. (The EEOC also may use binomial distribution (standard deviation) analyses to assess adverse impact in selected cases.)

Because alternative testing procedures without adverse impact often are scarce or nonexistent, test users must rely on documentation that their tests are job related. Job relatedness requires the test user to validate the test for its intended purpose in accordance with professional measurement standards. The test must measure bona fide occupational skills that have been demonstrated to be necessary for success on the job.

If Title VII had applied to the insurance licensure exams in the *Golden Rule* case, the company challenging the tests probably could have established adverse impact. The burden then would have been on ETS and the Illinois Insurance Department to demonstrate that the tested skills were job related. This requirement could be satisfied by documenting the linkage between the content of the test and the knowledge component of the entry-level insurance agent's job that is judged to be critical or important for a successful licensure applicant to demonstrate. If the job relatedness requirement were satisfied, the court would permit continued use of the test as long as the complainants could not demonstrate the availability of less discriminatory alternatives having substantially the same validity.

The lawsuit was settled out of court, the case was filed in a state court, and there have been *no* subsequent cases with similar facts. It is not certain at this time whether federal courts could apply Title VII employment requirements to state licensure tests, although several courts have ruled that Title VII does not apply to licensure testing. However, licensure tests still must satisfy the APA/AERA/NCME *Standards for Educational and Psychological Testing* (*Standards for Testing*). While acknowledging the inevitability of its use in litigation, the authors of the *Standards for Testing* emphasized the importance of professional judgment in determining its relevance and applicability in particular contexts.

Equal Protection—Proving Intent

Equal protection claims under the Fourteenth Amendment of the U.S. Constitution require (1) state action; (2) two classes, one of which is disadvantaged relative to the other; (3) classes based on race to receive a stringent standard of review; and (4) evidence of intent to discriminate. When membership in the class allegedly being discriminated against is based on race, the court will uphold the challenged government action only if it finds a compelling government interest and narrowly tailored means to achieve that interest. A compelling government interest is a very important reason for treating a racial group differently. Narrowly tailored means limit the government action to the absolute minimum action necessary to achieve the government's goals. However, when a government action is reviewed under a stringent standard, it is almost impossible for the government to meet this heavy burden; the challenged practice usually is found to be unconstitutional.

State action that disadvantages persons based on characteristics other than race still can be challenged under equal protection, but the odds are high that the challenged practice will be found constitutional. The lenient standard of review applicable to such cases places only a nominal burden on the state to show that it has a legitimate reason for its actions. For example, given a plausible justification for its actions, the state may treat persons differently based on wealth and have its actions upheld by the courts.

Therefore, in testing applications, challenges based on equal protection are likely to succeed only when the disadvantaged group is a protected racial group such as African-Americans, Hispanics, or Native Americans. Gender discrimination is reviewed under a standard between the stringent standard used for racial groups and the lenient standard used for socioeconomic, age, and other nonprotected characteristics. Thus, a challenge based on gender is more likely to succeed than one based on poverty, but less likely to succeed than one based on race.

The only other ways to mount a successful challenge when a stringent standard of review does not apply to the disadvantaged group are to demonstrate to the court's satisfaction that the state is using arbitrary, capricious, or irrational means to achieve its goals or that a fundamental right of a nonprotected group is being violated. Yet, few rights are viewed as fundamental by the courts. Freedom of expression and the free exercise of religion are fundamental rights, but education and employment are not.

Assuming that one can overcome the obstacles related to the type of group being disadvantaged, establishing the intent requirement of an equal protection challenge is still a formidable hurdle. Intent can be demonstrated by the inclusion of discriminatory language in the challenged statute or by discriminatory facts and circumstances surrounding the enforcement of a statute that appears neutral on its face. One relevant fact is the existence and magnitude of adverse impact on the disadvantaged group, but the courts have ruled that adverse impact alone is not enough to satisfy the intent requirement. The challenger must produce additional evidence of discriminatory intent, including but not limited to such factors as legislative history, application procedures, foreseeability/inevitability of disparate impact, sequence of events, deviation from normal procedures, and historical background. Effectively, the challenger must show that the state acted deliberately to *cause* adverse effects on a disadvantaged group, not that those adverse effects occurred *in spite of* the state's action.

In the *Golden Rule* case, the company had adverse impact statistics for a protected group. But the company would have needed more than these statistics to establish an equal protection violation. Because the statute authorizing the licensure testing was neutral on its face, the company would have needed to produce some additional evidence that ETS or the Illinois Insurance Department had developed and implemented the licensure tests in a manner designed to cause African-American applicants to fail.

The company alleged that it would prove intent to discriminate by showing that the tests did not comply with external professional standards, that no pretest or administration data were collected separately for historically disadvantaged groups, that historically disadvantaged group members did not participate in the test development process, that no job analysis was completed until after initial testing, and that ETS and the Department knew or should have known of substantial group differences in licensure test performance and took no corrective action. ETS and the Department disputed the company's ability to produce proof of these alleged indicators of discriminatory intent. Because the case never went to trial on the merits, we do not know what the evidence would have shown or how the court would have ruled.

However, in a more recent equal protection challenge to an admissions test required for entry into a teacher education program, the court ruled that knowledge by the decision-maker that a test will have an adverse impact on a protected group is not sufficient to establish intent to discriminate against that group. Because group performance differences are a combination of achievement differences and potential bias, test developers can be held accountable only for eliminating a portion of such differences by removing any demonstrable bias. The elimination of group differences remaining after bias has been removed would require remediation of the root causes in the educational and social environment.

Fundamental Fairness Under Substantive Due Process

Even when intentional discrimination against a protected group cannot be proven and the right being denied is a nonfundamental employment right, the state must still satisfy the fundamental fairness requirement. The fundamental fairness requirement derives from the substantive due process clause of the Fourteenth Amendment. Fundamental fairness means that the state action must not be arbitrary, capricious, or irrational. In testing applications, the courts have interpreted fundamental fairness to require that the test be valid as defined by the APA/AERA/NCME *Standards for Testing*. In addition to general validity evidence for employment testing, several courts also have required the test user to present a job analysis as part of its validity evidence. Generally, a job analysis is data from experts in the field and job incumbents that describe the frequency of use of various job functions, knowledge, and skills, and whether they are critical to the job.

In the *Golden Rule* case, the company alleged that the licensure test requirement was arbitrary because the skills tested had no relationship to what an insurance agent needed to know. ETS and the Department were prepared to refute this claim based on analyses of entry-level insurance agents' and brokers' jobs, which had been conducted and documented.

According to ETS and the Department, the job analyses, which were linked to the tests, helped demonstrate that the test items measured skills that minimally competent insurance agents and brokers needed to have. Significantly, the eventual settlement agreement in the case contained no provision restricting the substantive content of the licensure tests.

The *Golden Rule* Settlement

A settlement is not a court order—it is an agreement between two parties to a lawsuit. In dismissing a lawsuit after settlement, the court simply acknowledges that the parties have settled their differences and that there are no longer any issues requiring judicial intervention. The court does not evaluate the content of the settlement and makes no ruling regarding it. Thus, a settlement is binding only on the parties who agree to it and provides no legal precedent for any other lawsuit in any other court.

Despite its lack of legal authority, a settlement with one entity may be used to pressure another to agree to the same terms, which is what happened with the *Golden Rule* lawsuit. The terms agreed to in the settlement of the case were used to pressure legislators and other testing agencies to adopt similar procedures in other contexts. For example, in *Allen v. State Board of Education*, the complainants used a ratcheted-up, more stringent version of the *Golden Rule* settlement to pressure the state into an ill-advised consent decree severely limiting the state's latitude in the use of a teacher certification test. These extensive attempts by critics of testing to use the *Golden Rule* settlement as a precedent, even though the legal system would not view it as such, have made the content of the settlement important to policymakers who may be told that following its terms will eradicate test discrimination by eliminating biased test items.

The *Golden Rule* lawsuit was settled out of court in 1984, after eight years of procedural litigation. The parties agreed to the settlement to avoid further litigation and its associated costs. Neither party made any admissions regarding the allegations in the case. The parties agreed that the terms of the settlement would remain in force for seven years following dismissal of the case.

The Golden Rule Life Insurance Company claimed victory and promised that the settlement would open the test development process to greater public scrutiny and would decrease differences in performance between African-Americans and Caucasians. ETS emphasized that the settlement would affect only two of the four insurance licensure exams administered by the Illinois Insurance Department. But FairTest, a Boston organization originally financed by the Golden Rule Life Insurance Company, began campaigning for the adoption of the settlement terms by testing programs in other states.

The national visibility of the *Golden Rule* settlement, together with continued claims by testing critics that it reduces discriminatory bias in testing, makes knowledge of its terms and associated disadvantages important to policymakers. The key terms of the *Golden Rule* settlement were as follows:

(1) All test items were to be classified into one of two categories based on performance differences between African-Americans and Caucasians. Type I items had correct answer percentages greater than 40% for African-Americans, Caucasians, and the total of all groups, *and* had African-American/majority percentage correct differences of less than 15%. All other items were Type II.

(2) Type II questions were to be used only when no Type I questions were available in a content category.

(3) Items with the smallest group differences were to be used first.

Other administrative provisions providing greater external oversight and public access to information also were included in the settlement agreement. However, it was the rules for selecting items that generated the greatest controversy and that were cited by critics as precedent for other testing programs.

Advantages of the Settlement Terms

The company announced that the settlement would remove unfair obstacles to the licensure of qualified African-Americans. The company also believed that the item selection procedures mandated in the settlement agreement would be easily understood by the lay public and could be applied with varying group difference criteria in other testing programs.

At the time of the settlement, ETS believed that the terms of the settlement would be relatively easy to implement and would have only a minimal impact on the development of the insurance licensure tests. The costs associated with compliance appeared to be significantly less than the anticipated costs and delays expected for continued litigation.

Disadvantages of the Settlement Terms

Although many African-Americans were pleased to see equity considered in the selection of test items, some questioned the appropriateness of mechanical application of arbitrary rules (Bond, 1987). Subsequent research suggested that adverse impact reduction techniques like those in the *Golden Rule* settlement change the distribution of item difficulty but do not negatively affect other psychometric properties of the test as long as the item pool is sufficiently large. But when the item pool is more limited, content representation, validity, and reliability of test forms may be negatively affected.

With respect to the extension of *Golden Rule* procedures to other testing programs, one researcher stated that it would be inappropriate for K-12 achievement testing. The researcher argued that no African-American or Caucasian parents would want to be misled into believing that their children had learned material that they actually had not mastered (Bond, 1987).

Despite these concerns, the *Golden Rule* procedures were adopted by the parties in an Alabama teacher testing case, *Allen v. Alabama State Board of Education*, using a more stringent group difference criterion of 5%. Items for which the difference in the percent of African-Americans and Caucasians answering correctly was less than 5% had to be given first preference when selecting items for the teacher licensure tests. Items with group differences greater than the *Golden Rule* criterion of 15% could not be used at all. That is, if 95% of Caucasians but only 79% of African-Americans could correctly answer an item, that item could not be included on the teacher licensure test. After further consideration, the state temporarily abandoned its current teacher testing program rather than implement this restrictive procedure.

In hindsight, three years after agreeing to the settlement, ETS declared that accepting a compromise in the *Golden Rule* case had been a mistake and urged that it not serve as a precedent for other testing programs (Anrig, 1987). ETS cited three reasons for this belief: (1) the settlement had been interpreted by testing critics as an admission of guilt; (2) compromise procedures narrowly tailored to one small testing program were being cited as precedent for legislation and to obtain leverage in litigation involving testing programs in other states; and (3) the *Golden Rule* procedures were at best cumbersome and crude equity indicators and had not achieved the goal of reducing the adverse impact of the insurance licensure exams.

Perhaps the most serious criticism of the *Golden Rule* procedure was that most measurement professionals discredited it. Dr. Lloyd Bond, a spokesperson for the minority perspective, stated, "The psychometric profession is virtually unanimous in the condemnation of the *Golden Rule* as a bad precedent" (Bond, 1987, p. 20). Although the *Standards for Testing* recommend that differential item performance be considered as one piece of data in the item selection process, most measurement experts view the crude procedures mandated in the *Golden Rule* settlement as lacking credibility and scientific rigor (Mehrens, 1987; Shepard, 1987; Bond, 1987; Jaeger, 1987; Linn & Drasgow, 1987; Resnick, 1987). As explained below, much better state-of-the-art statistical procedures exist for detecting differential item performance and eliminating potentially biased items without significantly diminishing the validity of the test.

Effects of the *Golden Rule* Remedy on Test Validity

The *Golden Rule* remedy has been uniformly criticized by psychometric experts as a bad procedure for identifying potential test item bias. Although most experts agree that differential item performance should be considered when test items are selected, they believe that the *Golden Rule* remedy is an arbitrary rule inconsistent with state-of-the-art professional practice. Dr. Shepard stated that following the *Golden Rule* procedure "will lead to an inverse selection of items in terms of their psychometric adequacy" (Shepard, 1987, p. 7). The *Golden Rule* remedy has three major psychometric problems: (1) differences in historically disadvantaged group/majority item performance do *not* indicate bias; (2) sets of items selected using these differences will be the least reliable and most prone to the effects of guessing; and (3) the content validity of the test will be negatively affected when some subskills have no Type I items (Shepard, 1987; Linn & Drasgow, 1987). These issues are considered in greater detail in the following sections (see Phillips, 1990, for a more extensive discussion of the claims and counterclaims regarding *Golden Rule* and alternative procedures).

Differential Item Performance Is Not Bias Per Se

Responsible test users and test developers clearly do not want tests to exploit cultural differences. If certain vocabulary is likely to mislead a particular group, more neutral alternatives should be sought. Similarly, the portrayal of protected groups in ways that are viewed as demeaning and condescending should be avoided. But absent any offensive language, tested knowledge and skills that are supported by content experts should be considered appropriate for inclusion on a test. If group 1 has received instruction on the skills tested and group 2 has not, there may be a curricular validity problem. However, the remedy is not to delete items that group 2 answers incorrectly more often than group 1, but to change the instruction for group 2 to provide appropriate preparation.

Researchers have identified environmental factors as the key elements in the observed performance differences between African-Americans and Caucasians (Linn & Drasgow, 1987; Anastasi, 1961). This conclusion is consistent with the design of achievement and licensure tests to measure current knowledge and skills, not innate ability. Because current knowledge and skills are influenced by environmental factors, it is not surprising that differential life experiences result in differential knowledge among groups.

The most salient environmental factors identified by researchers are parent education and family income. These variables are closely related to past educational opportunities and poverty. Therefore, if poverty and inferior educational experiences cause one group to have less knowledge than another, should the test be condemned as biased or should the underlying environmental factors be targeted?

Concealing the effects of environmental influences by altering the test will bury the undesirable message but will do nothing to correct the underlying causes of differential performance. According to Dr. Anne Anastasi, "No test score can eliminate causality. Nor can a test score, however derived, reveal the origin of the behavior it reflects. If certain environmental factors influence behavior, they will also influence those samples of behavior covered by tests" (Anastasi, 1961, p. 389). Drs. Linn and Drasgow add: "The elimination or artificial reduction of differences in average test scores might conceal their situation, but it would not rectify it" (Linn & Drasgow, 1987, p. 15).

Test vs. Item Bias

The foregoing sections discussed concerns about whether an entire test was culturally biased. Such concerns usually view the test as a whole and make judgments about its appropriateness for historically disadvantaged groups.

In addition to such global judgments, the issue of potential bias against historically disadvantaged groups can be viewed as a local concern specific to individual test items. Even when global judgments indicate that the test as a whole is measuring appropriate content, individual items may be offensive to or function differently in historically disadvantaged groups. Because one would expect performance differences between groups from different environments, differential item performance, per se, is *not* evidence of bias. Item bias occurs when members of two groups similarly situated perform differently on an item. Similarly situated means that members of both groups have similar abilities.

Identifying Biased Items

Theoretically, if one were searching for possible item discrimination, one would want the members of each compared group to be as similar as possible on variables relevant to achievement. As a practical matter, the only information typically available for measuring the ability of group members is their total test score. Thus, the relevant comparison in item performance is between persons in each group with the same total test score.

For example, on a 50-item test, one might compare the performance of African-Americans and Caucasians who had total test scores of 45-50, 40-44, 35-39, etc. If in each of these subgroups of similar ability Caucasians consistently performed better on an item than African-Americans, then one might hypothesize that the item was potentially biased against African-Americans.

Flagging items according to the procedure described above still does not prove bias. Test scores include measurement errors and group scores contain sampling errors. Thus, the items identified as potentially biased must be screened further to determine whether there is reasonable cause to believe them unacceptable. This followup screening is usually done by panels of content experts that include representatives of relevant historically disadvantaged groups. If such panels believe the item to be a fair measure of the tested skill and can find no flaw in the item or reasonable explanation for the differential group performance, then the item remains in the pool of available items. But if there is good reason to believe that the differential performance is a function of group membership rather than of differential knowledge, then the item is revised or discarded.

The important thing to remember is that the statistics do not determine bias—human judgment does. An item can be flagged by chance occurrence of unexpected results when the item really is not biased (Hoover, 1984). Similarly, an offensive item may not be flagged by the statistical procedure. Thus, all items selected for a high-stakes test should be carefully screened by review panels with members from historically disadvantaged groups.

To appreciate the ambiguity in differential item performance statistics, consider the fact that sometimes the bias is in favor of the historically disadvantaged group. That is, sometimes historically disadvantaged group members significantly outperform majority group members on an item and produce a statistic that suggests that the majority group is disadvantaged. Often such items are not flagged and do not receive any additional scrutiny. Effectively, people behave as if such bias is not real, and they ignore it. Thus, one must be careful not to overinterpret items flagged for potential bias against historically disadvantaged groups and to use sound judgment to determine what content is appropriate for testing.

The example given earlier for using total test score intervals to control for differences in ability is a crude approximation of the actual methods used to detect differential item performance. All methods accepted within the measurement profession compare performance of groups of equal ability (Shepard, Camilli, & Williams, 1985; Hills, 1989). But the methods do not always agree with respect to the items flagged.

The number of items flagged by a method depends on the stringency of the criteria used. The larger the differential in performance required for flagging, the fewer items will be flagged. But even when similar numbers of items are flagged, the methods do not necessarily flag the same items. However, if only the extreme outliers are considered, there is usually substantial agreement among methods and with traditional indicators of poor item quality.

Some researchers have criticized the use of the total test score as a surrogate measure of group ability or achievement. In most cases, no other measure is available or, if another measure such as supervisor ratings is available, it is less reliable than the total test score. Some critics also have suggested that group ability be measured by a set of nonbiased items. Although this technique has worked well in simulation studies, it cannot be implemented with actual tests, because no definitive criterion exists for determining which items are truly nonbiased. Total test performance remains the most accessible measure for ensuring that group item performance comparisons are based on groups of equal ability.

Reliability and Validity

In addition to the overall reliability and validity of a test, each individual item must be valid and reliable. A valid item measures the intended objective or cell in the test specifications. Valid items present a clear, unambiguous question to the examinee with specific directions as to how the examinee is expected to respond. Valid items are free from offensive language or cultural stereotypes and have a "best" answer on which content experts agree. Valid items are free from grammatical cues, do not provide clues to other test items, and measure skills in the manner intended by the test user. Reliable items evoke consistent responses over time and among items measuring the same skill.

Valid and reliable test items serve the purpose for which the test was constructed. The purpose of diploma tests is to ensure that high school graduates have attained specified skills. The purpose of licensure tests is to protect the public from practitioners who are not competent. Both functions accrue to a state under its police powers to safeguard the health, safety, and welfare of its citizens.

Because the purpose of licensure or diploma tests is to ensure that specified content and skills have been attained, such tests divide examinees into two groups based on a passing standard set by relevant content experts. The function of the test is to decide for each examinee whether that person's knowledge and skills satisfy minimal professional standards. Thus, there are two categories of test takers: those who are competent and those who are not. The most valid items for such tests are the ones that provide maximum information about the examinee's status relative to the passing standard. Items that represent trivial or peripheral knowledge do not provide much assistance in making such decisions. Similarly, unless they represent critical knowledge, items that virtually all candidates answer correctly or incorrectly provide too little information to be cost-effective. However, items that measure knowledge and skills that competent examinees normally possess but that incompetent examinees normally do *not* possess will provide the best information for deciding the status of each examinee.

In contrast, the purpose of employment examinations is different from that of diploma and licensure tests designed to protect the public. The purpose of an employment test is to rank order applicants so that the most competent applicant(s) may be selected for hire. Thus, employment tests classify examinees into multiple categories of competence rather than making a dichotomous decision of competent/not competent. Valid items for employment tests must be job related—that is, they must measure skills that the applicant will use on the job. But such items also must differentiate among levels of competence so that those who have greater job-related skills may be identified. Therefore, rather than focusing on a single passing standard, the difficulty of items for an employment test will be spread across the continuum of relevant knowledge.

The employer's purpose for testing is to maximize productivity. The selection of an employee from a pool of job applicants is clearly a different task from that of granting a license that allows the applicant to become a member of the pool of eligible applicants. On employment tests, the issue for historically disadvantaged groups is whether the test accurately predicts their level of competence.

On the other hand, for diploma and licensure tests, the issue is whether the test accurately categorizes examinees from historically disadvantaged groups as possessing or not possessing the required knowledge and skills. The appropriate technical analyses for making this determination for state diploma and licensure tests are content validity reviews and discriminant analyses, not the predictive correlations or regression analyses common in employment testing applications.

The EEOC *Uniform Guidelines* were designed specifically for employment testing. Therefore, given the difference in purpose, it is questionable whether they should apply to licensure testing. Whereas the *Standards for Testing* require a test user to establish the validity of all assessments used to make high-stakes decisions about individuals, the *Uniform Guidelines* require validation only in the case of employment tests shown to have adverse impact on members of protected groups. However, in applications under each set of validation requirements, one must rely on professional judgment in selecting and applying relevant standards.

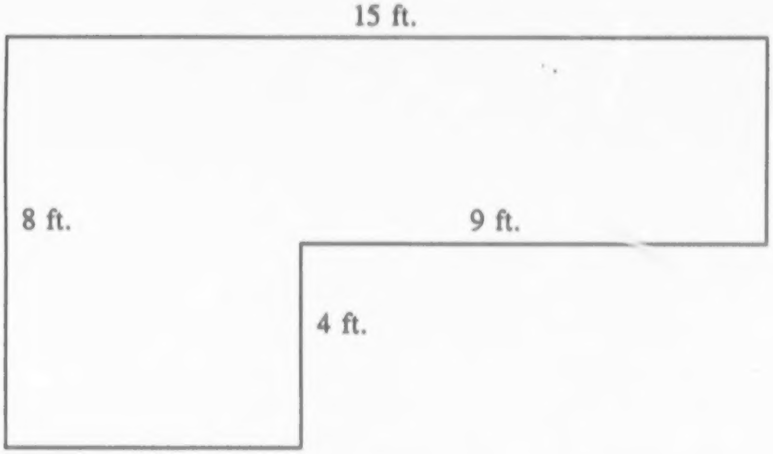
Neither set of standards specifies numerical criteria for judging item bias or item difficulty. Thus, the arbitrary numerical cutoffs for "acceptable" item difficulty and "tolerable" group differences in the *Golden Rule* settlement are not supported by professional standards. In addition, there is *no* connection between the arbitrary scope of the *Golden Rule* remedy and the alleged prior discrimination that the remedy was designed to correct.

Distortion of the Test Specifications

One of the more serious consequences of adoption of a *Golden Rule*-type procedure is the potential distortion of the test specifications (Linn & Drasgow, 1987). The test specifications detail the weighting of different content knowledge and skills and form the basis for item selection. For example, the test specifications for the mathematics section of a diploma test might call for 10% of the test to be devoted to solving word problems involving whole number and decimal computations. That is, if the mathematics test contains a total of 100 items, 10 items must be story problems.

A variety of questions might be developed to measure the ability to solve story problems. The following two items provide examples of easy and difficult story problems.

> 1. How many feet of fencing wire would Snoopy need to enclose a yard 50 feet wide and 100 feet long?
>
> 2. Garfield wants to carpet an L-shaped living room. The dimensions are given in the figure shown below. One wall contains a six-foot by eight-foot sliding glass door. If carpet costs $16.99 per square yard plus 4% sales tax, how much will it cost Garfield to carpet the living room?



Both items 1 and 2 are word problems involving whole numbers. But the solution to problem 2 requires more steps and more skills than problem 1. Problem 1 is a one-step problem involving only whole numbers. There are no changes of units and there is no extraneous information. The student need only add two lengths and two widths (100 + 100 + 50 + 50 = 300 feet).

But in problem 2, the student must do more reading, interpret a figure, sum two area computations, disregard the extraneous information about the door, change square feet to square yards, compute total cost with difficult decimals, convert a percent to a decimal to compute the sales tax, and add the tax to get the total cost. [(6 feet x 4 feet)+(15 feet x 4 feet) = 84 square feet/9 square feet per square yard = 9.33 square yards x $16.99 = $158.52 + .04($158.52) = $164.86].

Suppose problems 1 and 2 are given to a sample of tenth graders and the results are as follows for the majority group and historically disadvantaged group (HDG):

	Problem 1		Problem 2	
	majority	*HDG*	*majority*	*HDG*
Number	800	200	800	200
Percent Correct	92	84	63	41
Difference	8		22	

Problem 2 is more difficult than problem 1 for both majority and historically disadvantaged students. Using the *Golden Rule* procedure, both problems meet the 40% historically disadvantaged group percentage correct criterion. But problem 1, with an 8% difference in majority/historically disadvantaged group performance, would satisfy the *Golden Rule* differential performance criterion, while problem 2 would not. Problem 2 would not be acceptable, because its difference of 22% is greater than the 15% allowed by the *Golden Rule* procedure. Yet, problem 2 is much more realistic than problem 1 and involves the real world problem-solving skills that many math educators believe all high school students should master.

If all problems involving percents or extraneous information or more than one step had differential performance statistics similar to problem 2, and other one-step problems like problem 1 had acceptable differences, no problem 2 type items could be included on the test under the *Golden Rule* procedure. This requirement would severely limit the content validity of the test and would invalidate inferences to the domain of all story problems. One could only infer to a domain of one-step, whole number story problems with no extraneous information. Because differential performance tends to show up on more complex items, the test content could be seriously affected.

The above examples were made extreme to illustrate the effect that the *Golden Rule* procedure could have on content validity. But the concept generalizes to less obvious distortions of the test specifications. For example, in a cell labeled whole number computations, all long division problems might have differential performance statistics greater than 15%, eliminating all such items from the test. In spelling, all words of more than three syllables may have the offending statistics. But if these skills were part of the curriculum taught by all districts and they represented skills that educators judged important, it would be counterproductive to the goal of raising the level of competence of all high school graduates to eliminate them from the graduation test based on an arbitrary statistic. On the other hand, if there were a substantial difference in performance on problem 2 between majority and historically disadvantaged students for whom satisfactory algebra achievement had been independently demonstrated (e.g., who had received As or Bs in Algebra I), then a "cultural bias" explanation might be more plausible and the argument to eliminate the item more compelling.

When subareas of a content domain, such as multi-step story problems, are eliminated from the test, the *Golden Rule* procedure has effectively redefined the content domain. A major problem with such a redefinition is that it is impossible to separate that portion of the differential performance that represents real differences in learning (due, for example, to environmental factors) and that due to cultural discrimination. The *Golden Rule* procedure assumes that all performance differences are due to discrimination and seeks to minimize the overall difference in majority group and historically disadvantaged group means.

Two ways to counteract the distortion of content validity of a *Golden Rule*-type procedure are to develop a very large item bank and to use a set of content specifications broken down into specific subskills. Developing a large item bank makes it easier to find an item with the desired content that has acceptable differential performance statistics. Identifying particular content subskills in the test specifications means that multi-step story problems can be placed in a separate cell. If no items meet the *Golden Rule* criteria, then the least offending items may be used. The more detailed the content breakdown in the test specifications, the less likely that specific content will have to be left out of the test.

But making the content matrix more specific may have the disadvantage of limiting the sampling flexibility within the content domain of interest. Such specificity might preclude having three area and two perimeter items one year, followed by two area and three perimeter items the following year. The more detailed specifications might allow only two items of each type for each year's test.

On the other hand, the test developer might specify for a given year which cells in the content matrix will be included on the test. The next year, a different set of cells might be specified for testing. But once a set of cells has been specified for a given year's test, a Type I item must be selected for each cell, if available, and the test developer may not go to a neighboring cell in search of one.

Expert Testimony

The *Golden Rule* remedy was not adopted by the court, but it could have been adopted if the case had gone to trial. How could a court adopt a rule at odds with the majority of the measurement profession? The answer lies in the way courts receive and utilize expert testimony. Courts depend on expert testimony to establish facts in a case. Thus, the opinion of a single expert, if found credible by the court, can become the basis for a judicial finding of fact in a case.

Our legal system is adversarial. Each side makes the best case it can and the court arbitrates between the two opposing viewpoints. To serve the interest of the client, a lawyer will search for an expert whose opinions are most favorable to the client. So, for example, if 80 experts believe that X is true and only 20 believe that it is false, one expert of the group of 20 may be pitted against one of the 80. Thus, it appears that the opposing views are evenly balanced, even though the majority of the profession believes that X is true. The court may recognize and accord weight to the prevalent professional opinion within a scientific community. But having heard only one expert from each side, the court also will weigh the credibility, demeanor, and presentation of each expert. The more articulate expert who explains the issues in terms that the court can readily understand is more likely to have a favorable impact on the court as long as the views expressed are within reasonable professional boundaries.

When the court adopts a particular professional opinion in a case, it becomes influential in future cases. Thus, bad law can lead to bad science, as policymakers scramble to satisfy the dictates of the court. This possibility suggests the need for greater dialogue between the professional community and the legal community regarding appropriate standards. It also suggests that courts generally ought to refrain from imposing standards unless those standards fairly reflect prevailing professional opinion (which sometimes may be proven wrong later) and that matters of policy should be settled in forums other than the courts.

Reducing Group Differences

Differential item performance probably reflects past discrimination more than an intention to select tested skills that historically disadvantaged groups will fail to demonstrate. The test is merely the messenger of a serious social problem that will not be solved by ignoring majority/historically disadvantaged group performance differences with discredited techniques like the *Golden Rule* procedure. Improving the educational opportunities and expectations for disadvantaged groups is a more likely remedy.

Much debate has centered around the alleged narrowing of curricula caused by high-stakes tests, largely due to the inappropriate teaching of only those skills that are sampled on the test. If one also were to eliminate all skills that historically disadvantaged group members have not learned, curricula and licensure domains would be further narrowed. Can a state achieve its objective of protecting the public when it can require only knowledge shared equally by historically disadvantaged group and majority group members? For the substantial number of historically disadvantaged group members who do pass diploma or licensure tests, what does it mean to say that they have mastered skills beyond the expectations for their ethnic group?

Courts and testing programs must carefully weigh the short-term advantages of proposed new methodologies against the long-term welfare of historically disadvantaged groups and the public. Some would argue that giving away meaningless credentials will not result in real equality and will hide the social problems that so urgently need attention.

The real bias may be in differential instruction or limited life experiences rather than in the test item per se. If so, eliminating "discriminatory" items from a test will not solve the underlying problems that created the differential performance. Furthermore, such elimination may substantially reduce the content validity of the exam and the ability of the state to protect the public adequately from practitioners who have not mastered essential content.

However, one must always be vigilant for offensive language or cultural nuances that may inappropriately affect performance on an item. When screening items for potential bias, only techniques that compare groups of equal ability should be used. The *Golden Rule* procedure is not technically defensible under this criterion, since much better procedures are available. In addition, test developers should not rely on statistics alone; the final decision about the acceptability of an item should be made by historically disadvantaged group content reviewers. An items with no obvious biases should be discarded only when groups of equal ability perform differently on the item and experts from historically disadvantaged groups find the item unacceptable.

Extension of the *Golden Rule* Remedy to Other Applications

So far, professionals in the measurement community have successfully fought attempts in several states to legislate the *Golden Rule* procedure in statewide testing programs. The *Golden Rule* procedure has been introduced into proposed legislation in California, Massachusetts, New York, Texas, and Wisconsin. In several cases, different proposed cutoff values were substituted for the 40% and 15% values in the *Golden Rule* settlement.

In Texas, proponents of the *Golden Rule* procedure proposed that items on educational admissions and placement tests be required to have a 30% minimum correct answer rate for historically disadvantaged groups and majority/historically disadvantaged group performance differences no greater than 15%. Other proposed legislation in Texas would have mandated *Golden Rule* cutoffs of 40% and 10% for tests required for admission to teacher education programs in the state.

Legislation in California proposed that the *Golden Rule* procedure be expanded to include performance differences for five rather than two ethnic groups. Since Asian-Americans often out-perform all other groups, one might wonder whether that legislation would have required the elimination of all items for which the differential between Asian-Americans and Caucasians was greater than 15% in favor of Asian-Americans.

The most extreme attempt to apply the *Golden Rule* procedure occurred in the settlement of a teacher licensure testing case in Alabama, *Allen v. Alabama State Board of Education*. The lawsuit was filed by a class of African-American teachers in Alabama who alleged that they had failed the required licensure test because of discrimination. The settlement called for items to be separated into three categories. Priority was to be given to the first category for which the differential African-American/majority performance could be no more than 5%. If no items were available in the first category, items from the second category allowing 5% to 10% differences could be used. The third category included items with 10% to 15% African-American/majority percentage correct differences and were to be used only if *no* items from the other categories were available. Items with differences greater than 15% could not be used at all. Due to political maneuvering and procedural defects, the Alabama settlement was never implemented.

Summary

The courts have not ruled on the *Golden Rule* procedure. But critics of testing continue to push for its use in a variety of testing programs. Other testing cases suggest that courts will be unsympathetic to arguments about individual items and will judge the test as a whole. Courts will be interested in expert testimony about whether the test was developed using accepted and technically defensible measurement procedures that conform to appropriate professional standards. The courts also will be interested in whether experts from historically disadvantaged groups believe that the test measures appropriate content and will give some deference to the state's interest in protecting the public. States can minimize the likelihood that a court will impose a discredited remedy by actively seeking to implement appropriate methods for detecting and eliminating potential item bias.

Recommendations for Developing and Implementing Legally Defensible Item/Test Bias Review Procedures

(1) Establish a review panel of content experts representing all relevant historically disadvantaged groups (e.g., African-Americans, Hispanics, American Indians, females, and persons with disabilities) to review all items for possible offensive language, stereotypes, or cultural disadvantage prior to pretesting.

(2) When feasible, pretest all items before use. Alternatively, scrutinize all test items for bias after-the-fact and do not score items judged unacceptable.

(3) Calculate differential item performance statistics for relevant historically disadvantaged groups using a single professionally accepted method (e.g., item response theory or Mantel-Haenszel). Be sure that the procedure compares performance for groups of equal ability.

(4) Set criteria for flagging biased items to identify extreme outliers.

(5) Ask the review panel to re-examine all flagged items. If the panel as a whole and the historically disadvantaged members from the group for which the item was flagged

feel that the item is acceptable, retain it in the item pool or score it. If not, eliminate the item from scoring, revise it and re-pretest it, or discard it and write a new item.

(6) Monitor overall test performance for each relevant historically disadvantaged group. Identify areas of weakness by group and convey this information to educators or training programs providing remediation.

(7) Disseminate outlines of the content for which examinees may be tested. Provide clear explanations and examples of item formats, test administration conditions, and score interpretation.

(8) Involve members of relevant historically disadvantaged groups at all stages of the process, including selecting content areas to be tested; developing content specifications in each selected area; making policy decisions regarding item formats, testing time, security procedures, and accommodations; serving on item review and scoring panels; setting passing standards; reporting scores; and determining remediation procedures.

(9) Provide expert consultation to legislators who may be pressured by lobbyists to adopt inappropriate *Golden Rule*-type procedures.

(10) Use the media and public relations activities to inform the public and relevant constituencies of all activities and policy decisions related to the assessment program. Enlist their cooperation by providing clear rationales for each decision, seeking their input and answering their questions.

References

Cases and Statutes

Allen v. Alabama State Bd. of Educ., 612 F. Supp. 1046 (M.D. Ala. 1985), reh'g, 636 F. Supp. 64 (M.D. Ala. 1986), rev'd, 816 F.2d 575 (11th Cir. 1987).

Darks v. Cincinnati, 745 F.2d 1040 (6th Cir. 1984).

Georgia Ass'n of Educators v. Nix, 407 F. Supp. 112 (N.D. Ga. 1976).

Golden Rule Life Ins. Co., et al. v. Washburn, et al., No. 419-76 (Ill. 7th Jud.Cir. 1984).

Golden Rule Life Ins. Co., et al. v. Mathias, et al., 86 Ill. App. 3d 323, 408 N.E.2d 310 (1980).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Martin v. Educational Testing Serv., 431 A.2d 868 (1981).

National Org. for Women v. Waterfront Commission, 468 F. Supp. 317 (S.D.N.Y. 1979).

Schware v. Board of Examiners, 353 U.S. 232 (1957).

State v. Project Principle, Inc., 724 S.W.2d 387 (Tex. 1987).

Title VII of the Civil Rights Act, 42 U.S.C. §§ 2000e et seq. (1991).

Tyler v. Vickery, 517 F.2d 1089 (5th Cir. 1975).

Uniform Guidelines on Employee Selection Procedures (EEOC Uniform Guidelines), 29 C.F.R. §§ 1607 et seq. (1978).

United States v. City of Chicago, 549 F.2d 415 (7th Cir. 1977).

United States v. North Carolina, 400 F. Supp. 343 (E.D.N.C. 1975), vacated, 425 F. Supp. 789 (E.D.N.C. 1977).

United States v. South Carolina, 445 F. Supp. 1094 (D.S.C. 1977), aff'd mem, 434 U.S. 1026 (1978).

United States v. Texas, 628 F. Supp. 304 (E.D. Tex. 1985), rev'd on other grounds, United States v. LULAC, 793 F.2d 636 (5th Cir. 1986).

Wards Cove Packing Co. v. Atonio, 109 S. Ct. 2115 (1989).

Woodard v. Virginia Bd. of Bar Examiners, 420 F. Supp. 211 (E.D. Va. 1976), aff'd per curiam, 598 F.2d 1345 (4th Cir. 1979).

Articles and Other Resources

Anastasi (1961). Psychological Tests: Uses and Abuses. *Teachers College Record, 62*, 389.

Anrig (1987). ETS on "Golden Rule." *Educational Measurement: Issues and Practice*, *6*(3), 24.

Berk (1986). Judgmental and Statistical Item Analysis of Teacher Certification Tests, in Gorth & Chernoff (Eds.). *Testing for Teacher Certification*, 165.

Bond (1987). The Golden Rule Settlement: A Minority Perspective. *Educational Measurement: Issues and Practice, 6*(2), 18.

Ebel (1975). The Use of Tests in Educational Licensing, Employment, and Promotion. *Education and Urban Society, 8*, 19.

Green (1978). In Defense of Measurement. *American Psychologist*, *33*, 664.

Hills (1989). Screening for Potentially Biased Items in Testing Programs. *Educational Measurement: Issues and Practice*, *8*(4), 5.

Hoover (1984). The Reliability of Six Item Bias Indices, *Applied Psychological Measurement* *8*, 173.

Jaeger (1987). NCME Opposition to Proposed Golden Rule Legislation. *Educational Measurement: Issues and Practice*, *6*(2), 21.

Linn & Drasgow (1987). Implications of the Golden Rule Settlement for Test Construction. *Educational Measurement: Issues and Practice*, *6*(2), 13.

Mehrens (1987). Validity Issues in Teacher Licensure Tests. *Journal of Personnel Evaluation in Education*, 195.

Phillips (Dec. 20, 1990). The Golden Rule Remedy for Disparate Impact of Standardized Testing: Progress or Regress?, *Education Law Reporter*, *63*, 383.

Phillips (1991). Extending Teacher Licensure Testing: Have the Courts Applied the Wrong Validity Standard?. *T.M. Cooley Law Review*, *8*(3), 513.

Resnick (March 13, 1987). Letter on Behalf of AERA to Public Education Committee, Texas House of Representatives.

Rooney (1987). Golden Rule on "Golden Rule." *Educational Measurement: Issues and Practice*, *6*(2), 9.

Sales & O'Reilly (Jan. 1987). Scientific Judgments and the Law. *The Score*, 3.

Shepard (Jan. 1987). The Golden Rule Agreement: Bad Law, Bad Science, *The Score*, 7.

Shepard, Camilli, & Williams (1985). Validity of Approximation Techniques for Detecting Item Bias, Journal of Educational Measurement *22*, 77.

Thorndike (1982). *Applied Psychometrics*.

Weiss (1987). The Golden Rule Bias Reduction Principle: A Practical Reform. *Educational Measurement: Issues and Practice*, *6*(2), 23.

Chapter 4

Testing Accommodations for Persons with Disabilities

Overview

Traditionally, test administrators have provided testing accommodations for examinees with physical disabilities such as blindness or impaired mobility. Following passage of the Americans with Disabilities Act (ADA) in 1990, advocates for the disabled have argued that federal law also requires testing accommodations for cognitive disabilities such as dyslexia. But such accommodations may affect test validity, requiring policymakers to balance the social goal of integrating the disabled against the measurement goal of accurate test score interpretation. While the courts have provided some guidance regarding testing accommodation requirements for persons with disabilities, they have not yet addressed the issue of where to draw the line on accommodations for cognitive disabilities.

This chapter uses existing case law to construct a legal framework for considering accommodations for cognitive disabilities, explores the measurement problems associated with granting such accommodations, and discusses the advantages and disadvantages of alternative strategies for handling testing accommodation requests.

Terms

content validity
curricular validity
informed disclosure
job relatedness
liberty interest
otherwise qualified
predictive validity
property right
reasonable accommodation
reliability
substantial modification
undue hardship

Cases

Anderson v. Banks—denial of diplomas to special education students upheld because their disabilities precluded benefit from general education.

Board of Educ. of Northport-E. Northport v. Ambach—one year notice of a diploma testing requirement was not sufficient for special education students.

Brookhart v. Illinois State Bd. of Educ.—adequate notice and accommodations must be provided to disabled students subject to a diploma test, but substantial modifications of test content are not required.

Hawaii State Dept. of Educ.—The Office for Civil Rights (OCR) held that a reader must be provided for nonreading portions of state diploma test for learning disabled student and that accommodation requests must be decided on a case-by-case basis.

Southeastern Community College v. Davis—Supreme Court defined "otherwise qualified," holding that the college was not required to modify its nursing program to exempt a profoundly hearing impaired applicant from clinical training.

Legal Issues

Americans with Disabilities Act (ADA)
Education for All Handicapped Children Act (EHA)
Federal Legislation (ADA, EHA, IDEA, Section 504)
Fourteenth Amendment equal protection
Fourteenth Amendment procedural due process
Individuals with Disabilities Education Act (IDEA)
Office for Civil Rights (OCR)

Measurement/Educational Issues

"flagging" nonstandard test administrations
invalid accommodations
physical vs. cognitive disabilities
valid accommodations

Key Questions

(1) Why are accommodations for cognitive disabilities more problematic than those for physical disabilities?

(2) What are the characteristics of a valid accommodation?

(3) What are the legal standards for denying a requested test accommodation?

(4) What alternative policies are available to policymakers when responding to testing accommodation requests?

Concern for the treatment of disabled persons has become a national issue. The Americans with Disabilities Act (ADA) went into effect in 1992, requiring private entities to extend the same rights and accommodations to disabled persons as Section 504 of the Rehabilitation Act had required of public entities. Although a major provision of this legislation is to mandate the removal of physical barriers in building construction, it also prohibits discrimination against people with disabilities in employment and education. The regulations issued under the ADA expressly prohibit discrimination in testing.

Because the ADA was enacted only recently, case law has not yet been established under the Act. Section 504 cases suggest that the new legislation covers testing accommodations, but the courts have not indicated clearly which accommodations must be made under federal law and which may be denied.

Despite the lack of definition in the legislation and case law, test administrators are receiving increasing numbers of requests for variations in standard test administration conditions. These accommodation requests are coming from persons with a variety of disabilities who want accommodations such as a separate testing room, substantially more time to complete the test, frequent rest breaks, testing over several days, readers, scribes, sign language interpreters, transcriptions, desks and restrooms accessible to persons with limited mobility, typewriters, cassette recorders, large print or Braille booklets, magnifying equipment, calculators, computers with word processing/spellcheck/thesaurus, and oral/interactive test administration. Many requests for testing accommodations include multiple combinations of these options.

Physical vs. Cognitive Disabilities

It has been common practice to grant testing accommodations to persons with physical disabilities such as sensory deficits and mobility impairments. These commonly accommodated physical disabilities have included blindness, the use of a wheelchair, or a temporary incapacity such as a broken arm. Because the disability was obvious to anyone who interacted with the person requesting the accommodation, verification of the disability was not necessary. Moreover, the requested accommodations were clearly appropriate, because they primarily involved the removal of physical barriers and did not significantly affect the cognitive skills being tested.

For example, a common accommodation has been a Braille version of the test for the blind. Since it is obvious that loss of sight does not indicate impaired cognitive capacity, it makes sense to provide a person who cannot see the printed word with an alternative way to read the test questions. Additional time typically has been granted to a blind examinee as well, because reading in Braille is a slower process than reading printed materials.

Other common accommodations have involved access to the testing site. Ramps and elevators for wheelchair access, special restrooms that can accommodate wheelchairs and other physical apparatus, desks of appropriate height with removable chairs, and parking spaces close to the testing site have been given a high priority. Again, it is clear that accommodating the physical needs of the disabled provides a fair chance for them to take the same test that everyone else is taking. In general, few voiced concerns that this assistance would provide an unfair advantage or that the physical impairment cast doubt on the examinee's cognitive abilities.

One of the reasons for the lack of debate about accommodations for physical disabilities was that it was obvious when the disability itself disqualified the person for a particular activity. For example, even though a blind person could pass a written driving test in braille, everyone—including the person taking the test—would agree that becoming a bus driver was not possible. No one would dispute that sight is a requirement for driving.

More recently, however, test administrators have been receiving testing accommodation requests from persons with cognitive disabilities, such as attention deficit disorder, dyslexia, dysgraphia, and dyscalculia. In part, this increase in requests may be a function of increased diagnosis and treatment in elementary and secondary schools.

A generation ago, persons with such disabilities may have covered up their disabilities, dropped out of school, or even found themselves placed in institutions. But now that federal legislation requires states to provide appropriate elementary and secondary education for students with disabilities, these students are progressing farther in the educational system and are facing testing requirements for diplomas, college entrance, and professional licensure. Many educators believe that whatever accommodations disabled persons have received in their educational programs also should be made available during testing.

Accommodations for cognitive disabilities, however, can significantly affect the meaning and interpretation of the test score. The disability often is intertwined with the cognitive skills that the test user wishes to assess, and allowing the accommodation may effectively exempt the disabled person from demonstrating those skills. The test administrator then faces the policy dilemma of whether to allow a disabled person to substitute a different skill for the one measured by the test. In general, it is easier to demonstrate that a physical disability does not affect a tested skill than to demonstrate that a cognitive disability does not affect a tested skill. It is therefore more difficult to evaluate the appropriateness of substituted skills when the disability is cognitive and the skills tested also are cognitive.

The policy issue that decision-makers must address is whether accommodating a cognitive disability by providing a calculator for a math test, a reader for a reading test, or a word processor for a writing test is the same as accommodating a physical disability by providing wheelchair access or large print versions of a test. Ultimately, the debate centers around the apparent irrelevance of a physical disability to the skills being tested versus the perceived connection between a cognitive disability and these same skills. For example, one can imagine a paraplegic being an accomplished computer programmer, but it is more difficult to imagine a severe dyslexic being a successful journalist.

From a measurement point of view, the bottom line is whether the scores with and without accommodations are comparable. That is, do scores from nonstandard test administrations have the same meaning as scores from standard test administrations. This chapter will examine these issues more closely after providing an overview of applicable federal law and relevant court decisions.

Overview of Federal Statutory Requirements

Three major federal statutes have specific provisions for the disabled: the Individuals with Disabilities Education Act (IDEA), Section 504 of the 1973 Rehabilitation Act, and the Americans with Disabilities Act (ADA). Congress passed these statutes to correct serious abuses brought to its attention in testimony during hearings. For example, the IDEA was intended to provide educational services to disabled students who had been ignored, mistreated, or inappropriately institutionalized by the educational system. Section 504 addressed employment discrimination by public entities that refused to hire the disabled even when the disability was unrelated to the skills required for the job. The ADA extended this protection against employment discrimination to include private entities. Of particular concern to Congress in the ADA legislation was mandating barrier-free access to facilities open to the public.

Individuals with Disabilities Education Act (IDEA)

The Individuals with Disabilities Education Act (IDEA) is a 1990 revision of the Education for All Handicapped Children Act (EHA) enacted by Congress in 1975 to remedy prior failure of the public schools to educate disabled students. Rather than mandating substantive educational standards for disabled students, the IDEA relies on procedural safeguards to ensure appropriate educational services.

Specifically, the IDEA entitles a disabled student to the following: (1) a free appropriate public education; (2) an Individualized Educational Program (IEP), developed by a team of special education professionals, parents, and educational administrators, that describes in writing the disabled student's abilities and needs, the educational goals for the student, the specific services to be provided to meet those needs, and methods for evaluating progress; (3) related services such as transportation and support services (e.g., speech or physical therapy) that are necessary for the disabled student to benefit from special education; (4) educational services provided in the least restrictive environment appropriate to meet the disabled student's needs (i.e., integration with general education peers whenever possible, such as lunch, recess, physical education, music, etc.); and (5) procedures for parents to appeal any decisions with which they disagree.

While the IDEA clearly mandates specialized and individualized education for disabled students, the federal courts have held that it does not guarantee any particular educational outcome. That is, special education students are entitled to free educational services that meet their needs in the least restrictive environment, but they are not entitled to a high school diploma. A disabled student who has received appropriate educational services in an IEP but is unable to master the skills tested on a graduation test may be denied a high school diploma without violating the IDEA. However, federal regulations do require "good faith" efforts by the educational agency.

Section 504 of the Rehabilitation Act

The portion of Section 504 of the Rehabilitation Act that is relevant to testing persons with disabilities provides as follows:

> No *otherwise qualified* handicapped individual . . . shall, solely by reason of his handicap, be excluded from the participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance . . . .

Section 504 regulations issued by the Office for Civil Rights (OCR) further provide:

> A recipient [of Federal Funds] shall make *reasonable accommodation* to the *known* physical or cognitive limitations of an *otherwise qualified* handicapped applicant or employee unless the recipient can demonstrate that the accommodation would impose an *undue hardship* on the operation of its program.

Section 504 covers both physical and cognitive disabilities that affect one or more major life activities. It also covers any person who has a record of such an impairment or who is regarded as having such an impairment.

Physical and cognitive disabilities are broadly defined and include but are not limited to visual, speech, and hearing impairments; orthopedic impairments; cosmetic disfigurement; anatomical loss; muscular dystrophy; cerebral palsy; multiple sclerosis; epilepsy; heart disease; cancer; diabetes; cognitive retardation; organic brain syndrome; emotional illness; specific learning disabilities; drug addiction; and possibly alcoholism. Major life activities include walking, breathing, speaking, using the senses, performing manual tasks, caring for oneself, learning, and working.

In *Southeastern Community College v. Davis*, the Supreme Court defined "otherwise qualified" as a person who, despite the disability, can meet all educational or employment requirements. The Court held that the college was not required to modify its nursing program to exempt a profoundly hearing impaired applicant from clinical training. The Court was persuaded that the applicant was not otherwise qualified, because she would be unable to communicate effectively with all patients, might misunderstand a doctor's verbal commands in an emergency when time is of the essence, and would not be able to function in a surgical environment where required facial masks would make lip reading impossible. The *Davis* decision clearly indicated that an educational institution is not required to lower or substantially modify its standards to accommodate a disabled person, nor is it required to disregard the disability when evaluating a person's fitness for a particular educational program.

The meaning of the term "otherwise qualified" was further explained by a federal court in *Anderson v. Banks*. In the *Anderson* case, cognitively retarded students in a Georgia school district, who had not been taught the skills tested in a mandatory graduation test, were denied diplomas. The court held that when the disability is extraneous to the skills tested, the person is otherwise qualified; but when the disability itself prevents the person from demonstrating the required skills, the person is not otherwise qualified. Using this definition, the *Anderson* court reasoned that the special education students who had been denied diplomas were unable to benefit from general education because of their disabilities. The court further reasoned that the students' inability to meet academic standards for receipt of a diploma should not prevent the district from establishing such standards. The fact that such standards had an adverse impact on the disabled did not render the diploma test unlawful, in the court's view.

In a recent teacher licensure testing challenge, *Pandazides v. Virginia Board of Education,* an allegedly learning disabled candidate was denied individual, oral administration of the licensure test. However, the candidate's requests for additional time and transcripts of the audio portions of the test were granted. The federal district court upheld the state's denial of the request for individual, oral administration of the test. The case has been appealed.

Americans with Disabilities Act (ADA)

The Americans with Disabilities Act (ADA) was enacted in 1990 and became effective in stages in 1992. The ADA made minor changes in the wording of Section 504 and extended its public entity provisions to private entities. The major wording difference between the ADA and Section 504 that is relevant to testing was the substitution of "qualified individual with a disability" for "otherwise qualified handicapped individual." The revised language appears to be a cosmetic change, and its interpretation is likely to be the same as in cases decided under Section 504.

Although it is not clear that Congress intended to change prior testing law decisions, some disabled advocates believe that their rights under the ADA will be expanded. Future court cases and OCR rulings will be needed to clarify whether the standards for evaluating requested testing accommodations have changed under the ADA. In the interim—and without knowing exactly where the courts will draw the line between appropriate and inappropriate accommodations—policymakers will have to balance the social policy goal of including the disabled to the maximum extent possible against the measurement goal of obtaining valid test scores with consistent and meaningful interpretations for all examinees.

Constitutional Requirements: Equal Protection

Beyond attempts to enforce the specific federal statutory rights provided in the IDEA, Section 504, and the ADA, disabled persons could challenge alleged violations of their rights under the equal protection requirement of the Fourteenth Amendment to the U.S. Constitution. Recall that equal protection requires government entities to treat similarly situated individuals the same. However, the Supreme Court has reserved the highest standard of review for differential treatment of racial groups, and so far the disabled have not been designated a protected group for which the highest standard of review would apply.

Thus, an equal protection challenge based on differential treatment of persons with disabilities would probably receive low-level review, which would place the burden on the disabled to demonstrate that the government had acted arbitrarily or irrationally. Given the strong language of the federal statutes and the likelihood that the disabled would not be considered similarly situated in all circumstances, the disabled are probably more likely to be successful on a statutory challenge than on an equal protection challenge. This possibility did not deter the challengers in *Brookhart v. Illinois State Board of Education*, discussed below, from attempting to claim an equal protection violation.

Constitutional Requirements: Due Process

In addition to the federal statutes pertaining to the disabled, constitutional due process requirements apply. The Fourteenth Amendment provides that a state actor (government entity) may not deprive a person of a property or liberty interest without appropriate procedural safeguards and that the basis on which a property or liberty right is denied must be fundamentally fair. These safeguards against arbitrary government action that interferes with a property right are known as procedural due process and substantive due process.

As indicated in chapter 3, the court in the *Debra P.* case held that a diploma is a property right subject to Fourteenth Amendment protections. Implicit in that finding was the notion that actions by a state education agency or local school district qualify as state actions for purposes of the Fourteenth Amendment. Thus, any actions taken by a state education agency or local school district to deny a testing accommodation for a disabled person could qualify as the deprivation of a property right by a state actor if such actions result in a failure of the test and consequent denial of a diploma.

According to the *Debra P.* court, procedural due process requires adequate notice of the testing requirement. In that same context, the substantive due process requirement of fundamental fairness requires that the test have curricular validity. If these constitutional requirements are met, the *Debra P.* court held, the state may deny diplomas to members of a protected racial group (e.g., African-American students) who fail the state test.

The *Debra P.* case dealt with the general school population and did not specifically address the rights of disabled students. Extending the *Debra P.* holding to disabled students generates three major questions: (1) Can diplomas be withheld from disabled students who satisfactorily complete their IEPs but are unable to pass the graduation test? (2) If diplomas can be denied to disabled students, are the procedural and substantive due process requirements different in any way from the constitutional requirements for nondisabled students? and (3) What are the criteria for determining the testing accommodations that must be provided to disabled students, if any? The pre-ADA *Brookhart* case, litigated in a federal court in Illinois, provides some guidance in answering these questions.

Withholding Diplomas: The *Brookhart* Case

The *Brookhart* case addressed the procedural and substantive due process requirements for diploma tests applied to disabled students. This case involved a minimum competency test mandated by a local school district. All students receiving high school diplomas in this district were required to pass the minimum competency test. Students who failed the graduation test received certificates of completion but were denied diplomas.

The test covered reading, language arts and mathematics and the passing standard for each part of the test had been set at 70%. The testing requirement was imposed in spring 1978 and became effective for the spring 1980 graduating class. The graduation test was administered once each semester and students who failed were allowed to retake it until they passed or reached age 21.

Several disabled students who had successfully completed their IEPs but who had failed the graduation test and been denied diplomas filed a lawsuit to challenge the testing requirement and force school administrators to award them diplomas. A variety of disabilities were represented among the students challenging the testing requirement, including physical disabilities, multiple impairments, mild cognitive retardation, and learning disabilities.

In general, courts tend to be deferential to academic decisions as long as proper procedural safeguards are followed. In keeping with such deference, the *Brookhart* court held that disabled students could be required to pass a graduation test prior to receiving a high school diploma. However, as applied to disabled students, the *Brookhart* court modified the due process requirements of notice and curricular validity imposed by the *Debra P.* court.

Specifically, when tests are initiated as a requirement for a diploma, the *Brookhart* court stated that parents and educators must have adequate time to consider the disabled student's IEP and decide whether the tested skills should become part of the student's educational plan. Because the IEP process takes time and because disabled students may need more time to master tested skills, the court said that a longer notice period may be required for disabled students than for nondisabled students.

At minimum, the court held that less than 1½ years was not adequate notice for disabled students, particularly since this notice period may have been a slightly shorter period than that afforded nondisabled students. The court stated that incorporating the tested skills into disabled students' IEPs would take longer than integrating them into the general education curriculum. Together with the need for disabled students to be given more time than nondisabled students to master tested skills, the court was convinced that the disabled students who had been denied diplomas had not been given adequate notice of the requirement. However, the *Brookhart* court did indicate that parents and educators could decide that a student's IEP should not contain the tested skills, but only if the parents were given time to consider the consequences of receiving a certificate of completion rather than a diploma.

Another important holding in the *Brookhart* case was that test administrators are required under Section 504 to provide accommodations for disabled students. The court interpreted the Section 504 requirement of reasonable accommodations for a disabled person who is otherwise qualified to require physical accommodations such as Braille for the blind or wheelchair access. However, the court stopped short of mandating all requested accommodations. The court stated that a test administrator would not be required to grant an accommodation that "substantially modified" the test. For example, the court said that the test administrator would not be required to change the test questions.

In explaining the requirement for testing accommodations, the *Brookhart* court distinguished between factors in the test format or environment that prevented a disabled person from disclosing the degree of learning actually possessed and altering the test content because a person was unable to learn the tested skills due to a disability. According to the court, a person who is unable to learn because of a disability is not otherwise qualified, and the content changes necessary for such a person to pass the test would constitute substantial modifications, which are not required by law. This language in the *Brookhart* opinion suggested that the federal courts may be willing to draw a line between appropriate format accommodations and inappropriate substantive testing accommodations.

But the *Brookhart* court left open the question of whether accommodations for cognitive disabilities such as a reader, calculator, or word processor must be granted. Thus, the courts have not indicated which testing accommodations for cognitive disabilities are required under the ADA or Section 504. Given the dearth of legal guidance in this area, policymakers who must decide which testing accommodations to grant and which to refuse should consider how those accommodations might affect the validity of the examination—i.e., will test scores for persons receiving accommodations have the same meaning as test scores for persons who take the test under standard conditions? But before considering how decisions on accommodations might affect validity, this chapter will review the contributions of two state cases to the law on testing for people with disabilities.

State Cases

In addition to the *Brookhart* case, two state cases have considered testing of disabled persons. One was a New York case with facts similar to *Brookhart* and the other was a ruling by the OCR concerning decisions by the Hawaii Department of Education.

Both of these state cases have less precedential value than the *Brookhart* case, since *Brookhart* was decided by a federal court of appeals and therefore applies to the multi-state area within its jurisdiction. Although the other federal courts of appeal are not formally bound by the *Brookhart* decision, the decision does provide strong guidance for future judicial decisions and certainly would be considered by the other courts.

The New York case, *Board of Education of Northport-East Northport v. Ambach*, applies only to the state of New York. Although it could be cited in a case in another state, courts in other states are not required to consider it and could easily dismiss it in making their decisions.

The ruling in the *Hawaii State Department of Education* case comes from the federal agency charged with enforcing civil rights laws in the education field, the OCR. It is binding relative to future proceedings of that agency and could be cited in litigation, but it is not binding on any court.

Furthermore, in generalizing from all three of these cases, *Brookhart*, *Ambach*, and the *Hawaii* decision, one must remember that they were decided before the ADA became effective, and future court rulings under the ADA might differ from the holdings of these courts.

The Ambach *Case*

The *Ambach* case was brought by the Commissioner of Education in New York against a local school district. The school district had awarded high school diplomas to two disabled students who had failed to pass the mandatory statewide competency tests in reading and mathematics. The two disabled students—one had a neurological disorder and the other was trainably mentally retarded—had successfully completed their respective IEPs. The Commissioner sought to invalidate the diplomas awarded to these disabled students and any others who had not passed the required tests.

On appeal, the *Ambach* court found no Section 504 violation. The court stated that although Section 504 might require a district to make a school building accessible to the disabled by constructing a wheelchair ramp, it does not guarantee that a disabled student will be able to achieve the academic proficiency required to receive a high school diploma. The *Ambach* court also found no equal protection or substantive due process violations because, under constitutional analyses, disabled students are not a protected class and education is not a fundamental right.

However, based on a due process violation, the appellate court allowed the disabled students in the *Ambach* case to keep their diplomas. The court ruled that although the testing requirement had been announced three years before it became effective, disabled students effectively had notice of less than one year. The court reached this conclusion because written state guidelines specifically subjecting disabled students to the diploma testing requirement were not issued until the year in which the testing requirement became effective. Consistent with the *Brookhart* case, the *Ambach* court held that approximately one year's notice of a statewide testing requirement did not satisfy the procedural due process notice requirement. However, the court stated that if the disabled students had been given the full three-year notice period afforded nondisabled students, then the notice would have been adequate for the tested skills to be included in the students' IEPs where appropriate.

The Hawaii *Decision*

The OCR ruling in the *Hawaii* case dealt specifically with a challenge to a denied testing accommodation. The Hawaii Department of Education had refused a parent's request that her learning disabled son be allowed a reader for the statewide graduation test. The student's learning disability involved a processing deficit that substantially affected writing.

Department policy allowed readers only for nonreading portions of the test for students with certified visual impairments. These students were required to take the reading portion of the exam in Braille. Exceptions could be made to allow readers for special education students on nonreading portions of the exam, but the department normally denied all such requests.

The OCR agreed that allowing a reader for the reading portion of the test would defeat the purpose of the test and that denying it would not be discriminatory. But the OCR did find that denying a reader on portions of the test that were not designed to measure reading competency constituted unlawful discrimination against those disabled persons who have difficulties processing written materials.

Although the OCR ruling appeared to require test administrators to provide readers for any nonreading subtest, a careful reading of the opinion suggests that the real issue in the case was due process. The OCR opinion went on to state that because the needs and abilities of disabled students vary greatly, even when they have the same general disability, Section 504 requires that accommodations be judged on a case-by-case basis. But due to a large number of requests, the Hawaii Superintendent of Education had issued a directive to staff to grant requests for readers from blind students only. Thus, OCR seemed more concerned with the procedural aspects of administrative decision-making than with predetermining the outcome of any individual testing accommodation request.

Significance of the Hawaii *Decision*

Many educators probably were pleased with the OCR ruling in the *Hawaii* case. Because many of the requested accommodations are available to disabled persons for everyday tasks, advocates argue, they should also be allowed on the test. Allowing the disabled to use readers or calculators may be a small price to pay to increase these students' self-esteem. Some would argue that nonstandard testing conditions are just different ways to achieve the same result.

However, others are concerned about the potential corruption of the testing enterprise and the subjective basis upon which the label "learning disabled" is affixed to individuals. These issues are at the heart of the debate over the appropriateness of accommodations for cognitive disabilities. These and other measurement issues are explored further in the following sections.

Measurement Issues

Validity

A valid test measures what its users intend it to measure. Validity is not an inherent property of a collection of items, but a function of the manner in which the test scores are used and interpreted. A particular test may be valid for one purpose but invalid for another. For example, a final exam appropriate for a first-year high school algebra course would not be valid for assigning grades in a university statistics course.

A valid test must also be reliable—that is, it must measure consistently. This requirement means that if a person took a similar set of items on a different day, the person's score would be close to the score obtained on the original test. Because of measurement error, the two scores would not be identical, but the more reliable the test, the smaller the measurement errors and the closer the two test scores would be.

Validity and reliability must be determined separately for each test, because it is possible to have one without the other. For example, suppose the items on the algebra test measured the right skills but were so ambiguous that some had more than one correct answer, some could be interpreted more than one way, and some were so confusing that knowledgeable students could not understand what was expected. Or suppose the test consisted entirely of open-ended problems and the instructor graded the answers differently depending on how neatly they were written, whether all work was shown, or how the instructor felt at the particular time the questions were graded. In both cases, the algebra test might have content validity because it measures the intended skills, but it would lack reliability.

Similarly, a scale that consistently registers a person's weight as being ten pounds heavier than the true weight would misweigh persons consistently, but it would lack validity because it did not register any person's true weight. Although one might expect a measurement error of a pound or two, a consistent but unknown ten-pound error would invalidate the obtained weights.

The courts have recognized the importance of test validity by making it a part of the fundamental fairness requirement of substantive due process. The courts also have recognized the APA/AERA/NCME *Standards for Educational and Psychological Testing* (*Standards for Testing*) and the EEOC *Uniform Guidelines on Employee Selection Procedures* (*Uniform Guidelines*), both of which emphasize the importance of obtaining evidence of test validity (the former, generally; the latter, only upon a showing of adverse impact on legally protected subgroups). For diploma tests, the courts have required both content and curricular validity; for employment tests, the courts have required users to demonstrate predictive or content validity and job relatedness of the skills tested. Content validity requires that each sampled item represent a skill from the domain of required knowledge/skills. Job relatedness requires that the tested domain comprise knowledge and skills that experts and job incumbents believe are important or frequently used on the job.

The particular questions included in a test are a sample from the domain of all questions on that topic that could have been asked. Because time constraints preclude asking all possible questions or testing all possible subskills, the test user can obtain information on only a small fraction of the domain of interest. For example, if bar examiners want to know whether applicants understand criminal law, they can't ask about every possible crime under every possible combination of circumstances. They must settle for choosing a representative sample of questions from the domain of criminal law. If the sample is chosen systematically and there are no breaches of test security, the bar examiners can generalize from the particular sample of questions to the larger domain of interest (criminal law).

A major problem in any testing endeavor is measuring the right skills. Sometimes, no simple or straightforward measurement for the skill the user wishes to test is available. For example, suppose teamwork with co-workers is an important skill for a particular job. No readily available test for this skill exists. The best the employer can do is look for indicators in interviews and references. But teamwork is in the eye of the beholder, and very often one's judgment of such a skill is colored by one's approval or disapproval of the person's appearance, viewpoint, ethnic origins, personal habits, or other factors not relevant to the job.

To avoid the invalidity and unreliability in such judgments, the employer may decide to test the substantive skills involved in the job under the theory that those who are competent are more likely to make positive contributions to the team effort. Unfortunately, substantive competence and teamwork are not perfectly correlated; one can have competence without teamwork and vice versa. So the employee can argue that the employer is measuring the wrong skills, and the employer's alternative to testing substantive skills is to use a selection procedure known to be invalid and unreliable. Although this problem can be partially offset by using multiple measurements, the employer can still face a potential challenge to measurements least related to job performance or ones that are most subject to contamination by personal biases.

In the specific area of testing persons with disabilities, the *Brookhart* court stated that Section 504 and the EHA require a test to be valid for its intended purpose. The court defined validity in this context to mean that the test is suited for its intended purpose and suited for the particular population being tested. Although the *Brookhart* court held that test administrators must provide reasonable accommodations to the disabled, it also warned that such accommodations must not subvert the purpose for testing. The court left open the question of where to draw the line between valid and invalid accommodations, whether federal law requires separate validation of tests for the disabled, and the modifications, if any, required by the recently enacted ADA.

Valid vs. Invalid Accommodations

A 1987 OCR ruling on a Section 504 challenge to the Georgia statewide graduation test, *Georgia State Dept. of Educ.*, addressed the issue of separate test validation for the disabled. In its ruling, the OCR stated that once a statewide test has been validated as an appropriate measure of skills required for graduation, further separate validation for particular disabled groups would imply either (1) that the content of the test should be different for the disabled or (2) that the disabled should be tested in a different manner. The OCR went on to explain that adjustment of the test content would be inappropriate because it would be a substantial modification of standards contrary to the requirements of Section 504. With respect to the second implication, the OCR found the state in compliance with the reasonable accommodations requirement because the state had provided test format and environment modifications for disabled students. In a 1990 ruling, *Texas Education Agency*, the OCR reiterated its support for statewide testing guidelines that provide individually determined modifications for disabled students. Although the state in this case allowed complete or partial exemption from the testing requirement under certain circumstances, the OCR did not indicate that such exemptions are mandated by federal law.

When judging the appropriateness of a particular accommodation, test administrators should consider its effect on the content validity of the inference to be made from the test score. The APA/AERA/NCME *Standards for Testing* provide the following guidance:

> [U]nless it has been demonstrated that the psychometric properties of a test . . . are not altered significantly by some modification, the claims made for the test . . . cannot be generalized to the modified version. . . . When tests are administered to people with handicapping conditions, particularly those handicaps that affect cognitive functioning, a relevant question is whether the modified test measures the same constructs. Do changes in the medium of expression affect cognitive functioning and the meaning of responses? (p. 78).

Typically, there are too few examinees with each specific disability to conduct separate validity studies. However, aggregating the performance of several similar disabilities may not be appropriate due to extreme variations in severity within a single disability and the occurrence of multiple disabilities in a single person. The *Hawaii* decision in particular seems to suggest that individual decisions must be made after considering each person's disability.

Although it may be difficult to generalize about specific disabilities, some global decisions can be made about the appropriateness of specific accommodations for a particular test. When considering whether a requested test accommodation is valid or invalid, the test user should carefully consider the purpose of the test, the skills intended to be measured, and the inference the test user wishes to make from the test score.

The Purpose of Testing

All tests are developed for a particular purpose. The purpose for testing determines the types of items to be included, the skills to be measured, and the length of the test. Diploma tests and licensure tests are designed to protect the public from examinees who have not mastered minimum skills.

Although a worthy goal may be to measure each minimum skill in isolation, in reality it may not be feasible. For example, to read a mathematics test aloud to each examinee at a speed comfortable for that person would be prohibitively expensive and time-consuming. Test developers have compromised by constructing written mathematics tests at a reading level below that of the persons being tested. Subject-matter experts also have argued that the ability to read and comprehend mathematical symbols is part of the skill being tested.

Reading the test aloud and providing substantially more time to a disabled student would appear to disadvantage nondisabled students who read slowly, have limited vocabularies, have difficulty interpreting symbols, respond slowly, suffer test anxiety, or have difficulty staying on task. If these low-achieving students were allowed a reader and additional time, they probably would achieve higher scores than if the test were administered to them under standard conditions. One might be concerned that a low-achieving student who does not qualify for an accommodation would have less opportunity to demonstrate maximum performance than a student who has been labeled learning disabled.

Alternatively, some advocates would argue that it is fair to alter testing conditions for disabled persons because the alterations compensate for a neurologic disorder. Compensating for a neurologic disorder places the disabled person on an equal footing with nondisabled peers who do not have such disorders. This line of reasoning appears to divide cognitive testing difficulties into two categories: those caused by neurologic disorders and those caused by psychological problems. Again, one might wonder why students with neurologic disorder should receive an accommodation while students who simply get extremely anxious when taking a test are denied an accommodation. Neither group may be able to demonstrate their skills fully on a test administered under standard conditions. One hypothesis for the tendency to favor neurologic disorder may be the erroneous assumption that individuals have more control over psychological problems than over neurologic problems. Such thinking may create line-drawing problems when the origins of particular disabling behaviors are in dispute.

Fairness notwithstanding, the score of any person who is tested under nonstandard conditions does not have the same meaning as the scores for persons tested under standard conditions. One can reasonably assume that the business community and the public at large do not want diplomas and licenses to have different meanings for different individuals.

For example, suppose a person with dyslexia is unable to read printed text but was able to pass a graduation test that was read aloud to the person. The person obtains a high school diploma and applies for a job as a warehouse clerk for a large distribution center. The job involves handling invoices containing lists of items to be sent to a customer. The clerk must read each item on the list, find it on labeled shelves in the warehouse, pack it in the shipping box, properly address the box, and place it in the proper holding area for the section of the country to which it is being shipped. Knowing that the person has a high school diploma and being aware that the required graduation test includes a reading test, the warehouse manager is unlikely to give a reading test or ask specifically about reading skills.

Once hired, the person will need to have a coworker or other assistant read the invoices aloud. The employer may feel that providing a reader slows the process significantly and results in paying the cost of two workers to get the output of less than one. The disabled person, on the other hand, may suffer loss of self-esteem and livelihood if fired by the employer or forced to quit because of an inability to keep up with other workers doing the same job. Even if the disabled person had limited reading ability, the work might be considerably slower and more errors might be made.

If the employee described in the example above were a visually impaired person rather than a person with dyslexia, the interpretation of an accommodated score might be less cause for concern. These situations are distinguishable on two major levels: (1) listening versus reading comprehension; and (2) the job requirement of sight.

First, if accommodations for visually impaired persons were limited to large print or Braille editions, one could argue that the skill being measured is still the one the test is intended to measure: reading comprehension. The purpose of the large print or Braille accommodation is not to change the cognitive skill being measured but to remove the effects of the unrelated physical disability of visual impairment. Reading the test aloud, however, would confound the accommodation of the lack of sight with a change in the measured skill. As in the example given above where the test was read aloud to the person with dyslexia, administration of the reading test in oral rather than written form substitutes measurement of the skill of listening comprehension for the intended skill of reading comprehension. Thus, a blind student who passes the reading test using a braille edition has demonstrated competence in the intended skill of reading comprehension while not being penalized for the unrelated physical impairment of lack of sight. However, the applicant with dyslexia for whom the test was read aloud has not demonstrated competence in reading comprehension, because the accommodation in this case is related to the cognitive skill intended to be measured.

Second, the warehouse job described above included both reading and sight in its job requirements. The graduation test required for the high school diploma was intended to address the reading requirement; its purpose was not to determine visual acuity. Thus, the warehouse manager would be expected to assess the sight requirement separately from the reading requirement. Although a blind applicant who passed a Braille edition of the graduation test would have satisfied the reading comprehension requirement, the sight requirement would not be satisfied. Conversely, the applicant with dyslexia described above would have met the sight requirement but not the reading requirement. The warehouse manager could be misled by a graduation test purporting to certify reading comprehension but that really measured listening comprehension for this applicant. The manager might erroneously believe that the applicant with dyslexia satisfied both the sight and reading job qualifications because he or she has a high school diploma.

On the other hand, some jobs might reasonably allow listening comprehension to be substituted for reading comprehension. For example, a person with dyslexia who has experience fixing cars might be able to work in a garage as an auto mechanic. Even if the person could not read the labels on parts boxes, the person probably could tell the correct part by pictures on the box or by sight. If the paperwork were handled by the manager, the person might be successful at the job.

The problem with testing for diplomas and licenses is that the state is certifying a broad array of minimal skills that could be used in a variety of ways. Thus, there is no guarantee that the person with a reading disability will seek out a nonreading job. For example, a police officer with a visual impairment who passed the state certification exam with a reader would be qualified to ride in a patrol car as well as work a desk job. Nothing in the certification would prevent the person from applying for any entry-level law enforcement position, and, once hired, the employee might argue that providing a reader or altering the job to eliminate driving responsibilities would be a reasonable accommodation required of the employer. But what happens when the person's driving partner gets wounded in a shootout and the disabled person is the only officer on the scene available to pursue the suspect who flees the scene in a vehicle? Similarly, a left-handed applicant with a physical impairment of the right hand could not be licensed by passing the firearms test with only the left hand if certification standards require applicants to be able to discharge a weapon with either hand.

Clearly, protecting the public requires testing all essential skills for all activities for which the diploma or license recipient qualifies. The interests of society may be better served by providing incentives for employers to hire persons with disabilities in appropriate jobs than to mislead them into hiring persons who do not actually have the skills certified by the diploma or license.

This line of argument is not intended to minimize the inappropriate actions of some employers who have blatantly and inappropriately discriminated against disabled persons. For example, an employer who rejects a computer programmer in a wheelchair in favor of a less qualified mobile applicant has acted unfairly. Such situations are the kinds of obvious abuses that the federal legislation was designed to correct. The challenge in interpreting that legislation will be to correct the abuses without destroying the purpose for testing or rendering the test scores meaningless. This issue is discussed more fully in the following sections.

Invalid Accommodations

Suppose that a mathematics test objective stated: "The student will be able to do long division computations with pencil and paper." This objective would be substantially altered if the examinee were given a calculator during testing. On the other hand, suppose the objective stated: "The student will be able to solve multi-step story problems that include extraneous information." For this objective, computation might be considered only incidental and measurement of the objective might be facilitated by use of a calculator.

For both examples given above, the key question is: What does the objective require? Some math educators believe that computation should be de-emphasized and that all students should be given access to and training in the use of calculators. While this may be a worthy instructional goal, it is irrelevant to the validity of a particular accommodation. To determine whether an accommodation affects test validity, one must examine the test objectives as they are written. If the objective requires paper-and-pencil computation, then providing calculators will alter the inference from the test score. If proponents of calculators want students to be tested differently, they must first convince the test user to rewrite the objectives.

When an objective calls for a particular skill, such as long division on paper, those who pass the test will be assumed to be able to demonstrate the skill exactly as described in the objective. The inference from the test score will be that the student can do long division problems on paper. If the objective does not provide for the use of calculators, it is irrelevant that some people believe calculator computations are just as good.

Test scores should have the same meaning for all examinees—they should indicate what the examinees can do. If one examinee can do the paper-and-pencil calculations but another requires a calculator, the two examinees cannot do the same things. The inference from the first examinee's test score is to a domain of paper-and-pencil computations, whereas the inference from the second examinee's test score is to a domain of computations on a calculator. While in some situations either skill would be acceptable, in other situations the process for obtaining answers is as important as the answer itself. If test objectives are to communicate accurately to test users, the skills tested must be measured as specified in the objective so that the inferences made from the test scores will be the same for everyone.

Similar issues are involved in licensure and employment testing. For example, if a job requires an employee to answer the telephone and respond to oral queries by customers, listening comprehension is an important skill. If an applicant with an auditory processing deficit requested a transcript of the audio portions of a listening comprehension test as an accommodation, the applicant would be substituting reading comprehension for listening comprehension. But in the context of the job described in the example, such a substitution would not be reasonable. When customers call a business, they expect to interact verbally with an employee. Requiring customers to record messages for later transcription for an employee to read and respond to probably would create many dissatisfied customers. Even the accommodation of slowing down or replaying the audio tape would be impractical in this situation. Although a customer may be willing to repeat a query once, perhaps a bit more slowly, customers who must talk much more slowly than normal or repeat their queries over and over to be understood will probably become dissatisfied customers. The key issues here are whether the test measures relevant job skills and whether a requested accommodation alters the skill being measured.

Valid Accommodations

Sometimes the format of the test questions is not critical to the inference that the test user wishes to make. For example, many college professors are primarily interested in the knowledge that a student has acquired during a course and are not concerned with the students' reading abilities, writing skills, or their ability to answer questions quickly. In such circumstances, the professor may be willing to allow a student to demonstrate knowledge in written or oral formats and may be willing to allow ample time for all students to respond to all test items. Here, the format in which the knowledge is demonstrated is incidental and the inference to the domain of knowledge is still valid.

To minimize the potential for an invalid inference, a test user might want to grant only those accommodations judged essential. This approach might place some responsibility on the disabled examinee to adapt as much as possible to standard testing conditions. For example, some persons with dyslexia who have difficulty reading can improve their reading skills by using specially colored lenses or learning techniques for focusing their eyes on the page. Use of such techniques may mean that the person will read more slowly than others and it may be appropriate to allow some extra testing time. However, having the person with dyslexia actually do the reading is a closer approximation to the tested skill than having the test read aloud. The closer the accommodation to standard test administration conditions, the more valid the inference from the test score.

Administrative Decision-Making

Not all examinees with disabilities will require accommodations. For example, a person in a wheelchair may be able to test with other examinees if the building, testing room, and restrooms are wheelchair accessible. In other cases, a disabled person may be unable to take the test without an accommodation.

There is a fine line between testing accommodations that are valid and those that are invalid. Administrators must consider the purpose for testing and the skills intended to be measured. The wording of specific test objectives in diploma testing and relevant job requirements in employment and licensure testing are critical. When considering requested departures from standard testing conditions, administrators should consider the following questions:

(1) Will format changes or alterations in testing conditions change the skill being measured?

(2) Will the scores of students tested under standard conditions have a different meaning than scores for examinees tested with the requested accommodation?

(3) Would nondisabled students benefit if allowed the same accommodation?

(4) Does the disabled examinee have any capability for adapting to standard test administration conditions?

(5) Is the disability evidence or testing accommodations policy based on procedures with doubtful validity and reliability?

Answering yes to any of these questions suggests that an accommodation is not appropriate. The final decision of whether to grant a requested testing accommodation will require the test administrator to balance the individual rights of the disabled requestor against the state's obligation to maintain the integrity of the credential being awarded. The goals of providing maximum participation in society for the disabled and maintaining the validity of the testing program may be at odds.

Balancing the competing interests of the individual and society may require compromise and cooperation. But to avoid litigation when in doubt, the state may want to err on the side of granting the requested accommodation whenever feasible.

Classification of Disabilities

One difficult area for state administrators is verifying the disability. The administrator must decide whether a particular individual has the claimed disability, what accommodations are required for that disability, and whether those accommodations are appropriate. This task is made more complicated by disagreement among experts about what constitutes a particular disability and which individuals have the disability.

This problem is particularly acute in the area of learning disabilities. For students, studies have shown that classification as learning disabled depends in large part on the method used to identify the disability, the availability of services in particular disability categories, and the perception by the parent(s) of the benefit of special education for that student. Evidence suggests that it is very difficult to distinguish low achievers or slow learners from learning disabled students and that learning disabilities can become a catch-all category for any student deemed to need special attention in a variety of areas.

Studies also have raised questions about the interpretation of scores for learning disabled students who receive testing accommodations. In general, learning disabilities interfere with the cognitive ability to do academic tasks. Thus, a common accommodation is to allow additional time. But a study of learning disabled students who were given more time to complete the Scholastic Aptitude Test indicated that the resulting test scores overpredicted freshman grades. These data suggest that speed of work may play a role in academic success.

Unfortunately, some examinees develop a learning disability after failing a high-stakes test. Research also suggests that administrators may want to require current confirmatory evaluations for individuals diagnosed as learning disabled early in life. For example, medical researchers at Yale University have collected data indicating that the severity of dyslexia may diminish over time with maturation or special training.

Administrators also may want to screen carefully for the specific impairment of a learning disabled examinee. An odd curiosity in the *Hawaii* case was the request for a reader for a student with a claimed writing disability. It would seem appropriate to match the accommodation to the disability rather than allowing any accommodation once a learning disability has been documented.

Disclosure of Accommodations

For political reasons or to avoid litigation, test administrators may decide to grant testing condition accommodations that invalidate the test. In such cases, some test administrators have sought to protect test users from making erroneous inferences by adding notations to score reports, transcripts, or licenses that document the conditions under which a passing score was obtained.

However, the existence of a disability is "personally identifiable information" that is confidential. To the extent that identification of a testing accommodation also identifies the disability, the ADA may be interpreted to disallow such disclosure without the permission of the disabled examinee.

Even if judged appropriate, testing accommodation notations may be problematic in some cases for other reasons. For example, in licensure testing, such notations may not adequately protect the public from relying on practitioners with limited skills. The public may not know about the accommodation(s) if they do not see or read the license. The same argument also may apply to diplomas that may never be seen by employers.

Self-Selection with Informed Disclosure

In cases where notations of departures from standard testing conditions are feasible and sensible, administrators could decide to allow any student to request any accommodation without proving a disability. Applicants requesting accommodations could be asked to provide written permission for disclosure of the accommodation(s), but not the disability. Students who request testing accommodations also could be asked to sign a statement confirming notification of the test administrator's intent to disclose the accommodation(s).

When people who use the test scores, diplomas, or licenses know the conditions under which they were obtained, they are better able to interpret the scores properly. Self-selection of accommodations with informed disclosure would get administrators out of the business of judging which disabling conditions should receive accommodations and whether a particular person requesting an accommodation is actually disabled.

Under the ADA (and Section 504), persons are disabled if they have a physical or cognitive impairment, a history of such impairment, or are regarded as having such an impairment. Thus, an individual who does not actually have a disability can qualify if he or she is regarded as having one, or a person with a history of having a disability may qualify even if he or she no longer has the disability. Therefore, the accommodations may not be the most appropriate for such an individual's actual physical or cognitive condition. In cases where the reported disability is not extraneous to the skill being assessed and the requested accommodation would impair test validity, a more satisfactory procedure might be self-selection of accommodations with informed disclosure. The responsibility for deciding what the test score means would be left to the test user.

If self-selection of accommodations with informed disclosure is adopted, there must be no doubt that the examinee was given adequate information to make an informed decision prior to test administration. The requestor must be aware of the available accommodations, must be informed that the test user will be apprised of any departures from standard testing conditions when scores are reported, and must be given an explanation of the advantages and disadvantages of electing to take the test under nonstandard conditions. Such an explanation might include the potential for test users to misuse the notification and the tendency for test users to place greater value on nonaccommodated scores.

Informed disclosure, also known as "flagging" accommodated scores, has many critics. Even when test administrators follow the suggestions given above, disclosure of testing accommodations may result in a legal challenge from those who believe that such information should always remain private. Many advocates for the disabled believe that such notations will be misused to discriminate even when the disabled person is actually qualified. The *Standards for Testing* summarize the competing arguments as follows:

> Many test developers have argued that reporting scores from nonstandard test administrations without special identification (often called "flagging" of test scores) violates professional principles, misleads test users, and perhaps even harms handicapped test takers whose scores do not accurately reflect their abilities. Handicapped people, on the other hand, have generally said that to identify their scores as resulting from nonstandard administrations and in so doing to identify them as handicapped is to deny them the opportunity to compete on the same grounds as nonhandicapped test takers, that is to treat them inequitably. (p. 78).

One might argue that if an accommodation is needed to pass a diploma or licensure test, it also will be needed on the job; therefore, the employer has a right to know about it and "flagging" the score is not a violation of privacy. The possibility of hiding a relevant disability is particularly problematic for cognitive disabilities, which are not always obvious. For example, should a person with dyslexia who cannot read a reading comprehension test be allowed to hide that disability by not disclosing that a passing score was obtained by using a reader and extra time? This concern must be balanced against the possibility of discrimination against a person with a physical disability when the job requires only cognitive abilities.

Summary

Legal and measurement analyses suggest similar conclusions regarding testing accommodations. Testing accommodations may not be automatically denied. Test administrators must evaluate each request carefully before making a decision. Format accommodations that do not change the nature of the skill being measured should be granted. Requests for accommodations that would invalidate the inference made from the test score should not be granted. Those requests in the grey area in between must balance individual rights against the interests of the public. Whenever a test is administered under nonstandard conditions, the *Standards for Testing* and the *Code of Fair Testing Practices* recommend caution in interpreting test scores.

Because prior cases focused on physical rather than cognitive disabilities and were decided before the ADA was passed, we do not know where the courts will draw the line on testing accommodations. The Interpretive Regulations for the ADA suggest that the purpose of accommodations is to compensate for disabilities that negatively affect test performance but do not interfere with successful job performance. While many physical disabilities fit this requirement, cognitive disabilities may not.

To enhance defensibility in the event of litigation, test administrators are advised to develop and disseminate written policies. They also may want to consider self-selection of accommodations with informed disclosure.

Recommendations for Developing and Implementing Legally Defensible Testing Accommodations Policies

States may choose to grant nearly all accommodation requests or only requests based on documented physical disabilities, or they may choose a course of action somewhere in between. Whichever course a state chooses, legal defensibility will be enhanced by the development of a detailed policy and written procedures for the consideration of all requests. Careful consideration must be given to both the ADA requirements and test validity. Such policies also must protect the due process rights of the disabled. The following are suggested guidelines:

(1) Provide all school districts, training programs, and applicants for licensure with written instructions for requesting accommodations. These materials may be sent only on request, provided that their availability is communicated clearly in brochures and application materials.

(2) Provide a standardized form for requesting accommodations and clear directions for returning the application and all supporting materials to the state agency by a specified deadline.

(3) Require the requestor to provide documentation of the disability by a licensed professional experienced in diagnosing and treating the requestor's disability. A description of the disability and explanation of the necessity for the specific accommodation(s) requested should be provided in a letter signed by the licensed professional. Relevant test results and/or a description of the procedures used to make the diagnosis also might be required. The licensed professional should certify that his or her opinions are based on an in-person evaluation of the candidate conducted within the previous calendar year. In questionable cases, the licensed professional might be asked to provide documentation of his or her qualifications as an expert (e.g., a vita or biographical summary of relevant training, experience, and

professional memberships, plus licenses or certifications). The requestor or licensed professional also might be asked to supply relevant medical records.

(4) Require the requestor to provide documentation of any accommodations that have been provided in the requestor's educational or training program. This documentation should describe specific accommodations in detail and indicate the circumstances and frequency with which they were provided.

(5) If scores obtained under nonstandard conditions will be "flagged" or limited licenses granted, notify requestors of this fact and ask them to sign a statement prior to testing that confirms that they have been so notified. When the requestor is a minor, the parent(s) or guardian(s) also should sign.

(6) Designate a single individual within the state agency to review and act on all requests for testing accommodations. This person may be assisted in borderline cases by the opinion of a qualified consultant.

(7) Review testing accommodation requests on an individual, case-by-case basis, applying previously developed written criteria. Because disabilities differ in severity and an individual may have more than one disability, individual consideration is necessary. However, individuals similarly situated should be treated similarly. The state agency should develop general guidelines for accommodating various disabilities, but should review each case on its merits before making a final decision.

(8) At the state level, collect data on accommodations for cognitive disabilities if their effects on test validity are questionable. Such data may assist in gradually developing policies on "where to draw the line" in this area.

(9) Provide an expedited review procedure at the state level for all denied accommodation requests. Complete records of the documentation submitted by the requestor, phone calls, supporting materials received from professionals, correspondence, and the basis for the denial should be made available for the review. The review may be conducted by the agency head or a designated, qualified, impartial, outside expert hired by the agency. A written decision should be provided to the requestor.

(10) Upon written request, provide a formal appeal procedure—including a hearing—for the requestor when the denial of his or her request is upheld in the review process. Such procedures should follow the rules for administrative hearings and should allow legal representation and the presentation of evidence by the requestor. A formal hearing is useful even when not mandated by law, because it may resolve the dispute and avoid prolonged and costly litigation.

(11) Under the Individuals with Disabilities Education Act (IDEA), Section 504, and the ADA, students probably cannot be asked to bear any of the additional costs of providing testing accommodations. In licensure contexts in which the examinees bear

the testing costs, reasonable additional costs for accommodations may be acceptable. Reasonable limitations of accommodations to specific testing dates and sites are probably acceptable.

(12) To ensure stability and consistency across changes in personnel, state agencies may want to codify testing accommodations policies in administrative rules or legislation. Such rules also might indicate that test proctors have the responsibility for supervising the accommodations and specify the consequences for failure to follow state agency directives regarding nonstandard testing conditions.

References

Cases and Statutes

Americans with Disabilities Act (ADA), Pub. L. No. 101-336, 42 U.S.C. §§ 12101 et seq. (1990).

Anderson v. Banks, 520 F. Supp. 472 (S.D. Ga. 1981).

Board of Educ. of Northport-E. Northport v. Ambach, 436 N.Y.S.2d 564 (1981), aff'd with mod., 458 N.Y.S.2d 680 (A.D. 1982), aff'd, 457 N.E.2d 775 (N.Y. 1983).

Brookhart v. Illinois State Bd. of Educ., 697 F.2d 179 (7th Cir. 1983).

Georgia State Dept. of Educ., EHLR 352: 480 (OCR, 1987).

Hawaii State Dept. of Educ., 17 EHLR 360 (OCR, 1990).

Individuals with Disabilities Education Act (IDEA), Pub. L. No. 102-119, 20 U.S.C. §§ 1400 et seq. (1991).

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal. 1979).

Pandazides v. Virginia Bd. of Educ., ___ F. Supp. ___ (E.D. Va. 1992) (on appeal).

Section 504 of the Rehabilitation Act, 29 U.S.C. §§ 701 et seq. (1973).

Southeastern Community College v. Davis, 442 U.S. 397 (1979).

Texas Educ. Agency, 16 EHLR 750 (OCR, 1990).

Articles and Other Resources

Clarizio & Phillips (1992). A Comparison of Severe Discrepancy Formulae: Implications for Policy Consultation. *Journal of Educational and Psychological Consultation, 3*(1), 55.

Clarizio & Phillips (1989). Defining Severe Discrepancy in the Diagnosis of Learning Disabilities: A Comparison of Methods. *Journal of School Psychology, 27,* 383.

Gerber (February 1991). Survey of Current Practices of Testing Handicapped Bar Applicants: The State Bar Examiners Two Years Later. *The Bar Examiner,* 43.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education.* Washington, DC: Author.

Phillips (in press). Testing Accommodations for Disabled Students. *Education Law Reporter*.

Ragosta (February 1991). Testing Bar Applicants with Learning Disabilities. *The Bar Examiner*, 11.

Shepard (1983). The Role of Measurement in Educational Policy: Lessons From the Identification of Learning Disabilities, *Educational Measurement: Issues and Practice*, 2, 4.

Smith (February 1991). NCBE Guidelines for Testing Disabled Applicants, *The Bar Examiner*, 28.

Willingham, et al. (1988). *Testing Handicapped People*.

Chapter 5

Legal Issues in Performance Assessment

Overview

The recent pressure for inclusion of performance tasks in statewide assessments raises a number of legal and measurement issues that require careful advance consideration by policymakers. Developing and administering legally-defensible, large-scale performance assessments is labor intensive and extremely expensive. Although claims of authenticity may appear to increase the validity of performance assessments, this initial public relations advantage may be offset later when serious measurement problems are encountered and the implementation of solutions recommended by experts is precluded by budget crises.

To assist policymakers in applying relevant legal standards, this chapter develops a legal framework for evaluating performance assessments based on case law from employment testing, teacher testing, and higher education. The extension of these legal decisions to statewide performance assessment programs suggests that professional standards and expert testimony will play a key role in future legal challenges. Major measurement issues likely to be addressed by experts evaluating performance assessments are discussed and illustrated with examples.

Terms

burden of proof
content validity
curricular validity
disparate impact
equating
face validity
job analysis
notice
objective assessment
predictive validity
subjective assessment
passing standard
pre-equating
property interest
standard error of measurement

Cases

Griggs v. Duke Power Co.—sufficient statistical evidence of an adverse impact of an objective employment test existed to establish Title VII discrimination.

Perez v. F.B.I.—Supreme Court held that nonstandard phone interviews used to assess Spanish language skills of Hispanic agents for undesirable assignments were discriminatory.

United States v. South Carolina—upheld validation of teacher licensure test against teacher preparation programs rather than measures of successful teaching.

Wards Cove Packing Co. v. Atonio—required an employer to meet a less stringent standard of producing evidence to justify its employment practices.

Watson v. Ft. Worth Bank & Trust—Supreme Court held that standards applicable to objective tests also apply to subjective assessments.

Legal Issues

EEOC *Uniform Guidelines*
Fourteenth Amendment equal protection
Fourteenth Amendment due process
1991 Civil Rights Act
Title VII of the Civil Rights Act of 1964

Measurement/Educational Issues

APA/AERA/NCME *Standards for Testing*
content sampling
equating and pre-equating test forms
errors of measurement
narrowing the curriculum
potential "bias"
reliability evidence
scorer reliability
setting passing standards
standardization
testing as a vehicle for curricular reform
test security
validity evidence

Key Questions

(1) What legal requirements for performance assessments have emerged from employment and other testing cases?

(2) What are the major measurement issues policymakers need to consider when implementing performance assessments?

(3) Why are performance assessments so labor intensive and expensive?

(4) What type of legal challenge to statewide performance assessments is most likely to occur in the near future?

(5) What steps can be taken to minimize potential legal challenges to performance assessment programs?

Testing affects the life of nearly everyone today. Tests are used to award diplomas, select applicants for college, place students in special programs, hire and promote employees, and license professionals. When multiple-choice tests were first developed, test users believed a great advancement in testing technology had been achieved. Recent reforms have sought to solve testing problems by bringing back the old performance assessments that the multiple-choice tests replaced.

Origins of the "Authentic Assessment" Movement

Performance assessment is not a new idea. In the Old Testament, the Gileadites asked all persons seeking to cross the Jordan River whether they were enemy Ephraimites. If the answer was "No," the traveler was asked to say "Shibboleth." Ephraimites could be identified by their inability to pronounce the "sh" sound; those who said "Sibboleth" were seized and killed.

The most recent reincarnation of performance assessment has its roots in the high-stakes testing of the past few decades. Tests have become accountability tools as students, teachers, and schools vie for scarce tax dollars. Publication of school-by-school rankings on statewide tests, the use of tests to award state funds and high school diplomas, and other public disclosure of high-stakes testing results have made the testing enterprise very visible and put extreme pressure on tests to serve multiple purposes simultaneously while keeping actual testing time to a minimum.

With minimal information being used to make a maximum number of high-stakes individual and group decisions, it is not surprising that critics, believing the process to be unfair, have challenged testing programs in the courts. But what is a bit surprising is the rhetoric by advocates of performance assessment that suggests that it can solve the problems inherent in high-stakes testing. Instead of attacking the high-stakes uses of tests, some critics have attacked the format of the items and have declared that multiple-choice items are at fault for testing misuse. They believe that if all multiple-choice items are replaced by performance assessments, examinees will be required to demonstrate complex higher order thinking skills more consistent with good classroom instruction and real world applications. This claimed advantage of realism has led advocates to refer to performance assessments as "authentic assessments." Some cognitive psychologists also believe "authentic assessments" are superior to traditional tests because process skills can be given equal or greater emphasis than obtaining the correct answer.

However, some measurement experts doubt that performance assessments can live up to the sweeping claims made by advocates. They cite several reasons why performance assessments alone cannot solve all our testing problems. For example, (1) some knowledge can be measured more efficiently with objective items; (2) skilled item writers can produce challenging objective items that also measure higher order thinking skills; (3) insufficient research has been completed to document the claimed advantages of performance assessment for all testing applications; (4) performance assessments have inadequate technical properties for making high-stakes individual decisions; (5) the significantly increased costs of performance assessment are disproportionate to incremental information gains; (6) scoring is more subjective and thus prone to greater errors of measurement; and (7) performance assessments are more suited to classroom instruction, where incorrect decisions can be adjusted with minimal injury to the student, than to one-shot, large-scale, high-stakes accountability applications.

As later sections indicate, there is no reason to believe that inappropriate testing practices, breaches in test security, narrowing of the curriculum, adverse impact on historically disadvantaged groups, requests for testing accommodations, measurement error, or equating problems will magically disappear when performance assessments are substituted for traditional multiple-choice tests. In fact, preliminary data from large-scale assessments are beginning to suggest that many of these issues in high-stakes testing have worsened with the introduction of performance assessments (Braun, 1993; Harp, 1993). If this conclusion is correct, legal challenges to testing programs may increase in the future. With limited resources and tight budgets, statewide testing programs will need to plan carefully to minimize potential litigation and to produce the documentary evidence necessary to defend high-stakes performance assessment programs in the event of a legal challenge.

Given the current rush to implement untested performance assessments for accountability, the inclination of protected groups to file suits when denied benefits, and the doubts of some measurement experts regarding the technical adequacy of performance assessments for high-stakes decisions, a legal challenge is likely in the near future. To provide some indication of how the courts may view such challenges, holdings from traditional testing cases and related employment testing cases will be reviewed. These legal principles will then be examined in the context of relevant measurement issues.

Historical Perspectives from Traditional Testing Cases

Based on prior litigation involving traditional multiple-choice tests, the most likely challenges to performance assessments will focus on adverse impact on protected groups. Members of historically disadvantaged groups who are denied whatever "benefits" accrue to "passing" members of majority groups will probably challenge the validity, fairness, and notice of the assessment requirement. Such constitutional challenges are more likely the higher the stakes and the greater the differential in historically disadvantaged group/majority passing rates.

Equal Protection

As in other contexts described earlier, equal protection challenges require state action that disadvantages one group relative to another. To obtain the highest standard of review most likely to find the testing program unconstitutional, the disadvantaged group must be a protected racial or ethnic group. In addition, the challenger must present evidence of disparate impact and intent to discriminate.

For purposes of equal protection analysis, disparate or adverse impact can be demonstrated if, on average, members of the historically disadvantaged group receive lower scores than majority group members or if the failure rate is significantly higher for historically disadvantaged group members. (Note that Title VII adverse impact in employment cases is demonstrated by the "four-fifths rule" or standard deviation analysis under the EEOC *Uniform Guidelines*.)

Some advocates of performance assessments have claimed they are fairer to historically disadvantaged groups because they are "authentic." But preliminary data from high-stakes applications indicate that the gap between historically disadvantaged groups and majority performance may be widening when performance assessments replace traditional tests (Beck, in press; Mehrens, 1992). If one believes that it is more difficult to construct an answer from scratch than to select a correct answer from a set of choices, it may not be surprising to find the majority/historically disadvantaged group performance differential increasing. Thus, it appears that challengers will have ample evidence of disparate impact on performance assessments.

To convince the court the testing program violates the constitution, challengers must also show that test users adopted performance assessments with an intent to discriminate. It is unlikely that any formal statements to this effect will have been made; in fact, the rhetoric may be just the opposite. Performance assessments may have been adopted with the specific intent to be fairer to all students, particularly historically disadvantaged groups. Thus, to prove intentional discrimination, challengers will have to present facts and circumstances tending to indicate that the unstated real purpose of the assessment program was to deny benefits to members of historically disadvantaged groups.

Factors that might contribute to an intentional discrimination argument include culture-specific tasks, unreliable or "biased" scoring, nongeneralizable content sampling, lack of historically disadvantaged group representation in the test development process, inferior preparation at institutions enrolling students predominately from historically disadvantaged groups, off-the-record racial slurs by policymakers, and prior knowledge of greater disparate impact for performance tasks than for multiple-choice items. For example, if videotaped performances made it possible for raters to identify the respondent's race, this irrelevant information might inappropriately affect the scores given by some raters. Even if this did not occur, allegations of such potential "biases" would be difficult to disprove, particularly when scores for historically disadvantaged group members were consistently lower than those for majority group members.

Based on the "narrowly tailored" requirement adopted by the Court in equal protection challenges or the civil rights option of presenting the court with an "equally effective but less discriminatory alternative," historically disadvantaged groups might also argue against performance assessments by contrasting their larger adverse impact with that of multiple-choice tests where the gap in majority/historically disadvantaged group performance has narrowed over the last decade (see *Debra P. v. Turlington*). Using such data, historically disadvantaged groups might claim that performance assessments are just a new way of widening the gap once again.

Due Process

Performance assessment challengers might also claim a notice or validity violation based on court holdings in the *Debra P.* case or a procedural violation in the handling of requests for nonstandard test administrations for persons with disabilities. The *Brookhart* case suggests that courts might require longer notice periods for special education students.

The *Debra P.* fundamental fairness requirement that diploma tests have curricular validity applies to all tests, even when allegations of discrimination have not been proven. The substantive due process fundamental fairness requirement also dictates that testing programs not be arbitrary or capricious, even when there is no fundamental right or protected group involved.

Some advocates of performance assessments have claimed that validity evidence need not be collected because "authentic" assessments are valid by definition. But courts have rarely accepted arguments of "face validity" and test users who skimp on the collection of validity data may face embarrassing and costly consequences in the event of litigation. For example, test administrations by untrained classroom teachers may vary substantially across the state. Evidence of such variations could support allegations of lack of task standardization, bias, or lack of content or curricular validity.

Legal Perspectives on Performance Assessment

So far, there have been no federal court cases dealing with the use of performance assessments in high-stakes, diploma testing programs. This may be because, except for writing, states so far have not denied diplomas based on performance assessments. Writing assessments have been generally well-received, in part because: (1) the public believes that students should be able to write well; (2) the public also believes that to measure writing skills, students should be asked to actually write essays; (3) writing prompts and scoring rubrics have been narrowly defined and communicated in detail to educators and the public; (4) raters have received extensive training in applying well-specified criteria and in resolving any scoring discrepancies; (5) single sample responses scored by a single rater have not been overinterpreted or oversold by testing staff; and (6) schools have significantly increased the number of opportunities for students to practice their writing skills. But as states expand performance assessment into other subject areas, they may find less consensus about what should be tested and how it should be measured. States may also find their already limited budgets severely strained by the volume of resources needed to construct and score defensible performance assessments in multiple subject areas.

Legal decisions that have addressed performance assessment have dealt with employment and higher education applications. Although these applications differ in significant ways from secondary education, they indicate the perspectives and the kinds of standards that federal courts are most likely to adopt in diploma or licensure challenges to performance assessments.

The case law reviewed in the following sections includes challenges to subjective promotions, hiring criteria, dismissal from a training program, revocation of a college degree, the use of phone interviews and ethnic origin to assign language-related work, and nonrenewal of teaching contracts. In addition to the substantive standards suggested by these employment and higher education cases, they also illustrate how courts assign burdens of proof and how standards evolve over time.

Legal Perspectives from Employment Cases

Subjective Employment Decisions

In a 1988 Title VII employment challenge, *Watson v. Fort Worth Bank & Trust*, the U.S. Supreme Court held that the standards applicable to objective tests also apply to subjective assessments. Subjective assessments in employment testing have similar characteristics to performance assessments in education. Subjective employment assessments include interviews, informal observations, supervisors' ratings, and other judgments of competence based on unspecified data.

In the *Watson* case, an African-American bank teller had been repeatedly denied a promotion to a head teller position. Each time the African-American teller applied for an opening, she was passed over by a Caucasian male supervisor who subjectively evaluated her as not qualified, in spite of a history of good job performance ratings. In each case, the opening was filled by a Caucasian male or Caucasian female. It appeared that no formal criteria, ratings, or checklists were used by the supervisors evaluating applicants and no attempt had been made to construct a job analysis or to compare systematically the evaluations of different supervisors.

Two types of Title VII challenges can be filed by a person who believes that an employment practice is discriminatory: disparate impact and disparate treatment. Disparate treatment challenges allege that a particular individual has been treated in a discriminatory manner, whereas disparate impact challenges allege that members of the protected group to which the complainant belongs have been discriminated against by the employer's policies.

Prior to the *Watson* case, the federal courts were divided on whether disparate impact claims could be brought against subjective hiring practices. In 1971, the Supreme Court in *Griggs v. Duke Power Co.* applied disparate impact analysis to an objective test used to evaluate job applicants for entry-level positions. In several subsequent decisions, the Supreme Court applied disparate treatment analysis to adverse employment decisions based on personal judgments but did not specifically address the issue of whether disparate impact analysis also could be applied to subjective decisions.

In the *Griggs* case, the company had required all new hires either to have a high school diploma or to pass an aptitude test. The applicants who were not hired claimed that the test was not job related and was being used as a pretext to exclude African-Americans. Under a disparate impact theory, the court held that statistical evidence demonstrating adverse impact on African-American applicants was sufficient to establish Title VII discrimination.

The 1988 *Watson* case established that employees also could challenge subjective hiring practices under a disparate impact theory. This ruling was important because disparate impact claims are generally easier to prove than disparate treatment claims. The complainant in a disparate treatment claim must prove that the employer intended to discriminate, but in a disparate impact claim the complainant need only show statistically that the protected group has been significantly disadvantaged. In ruling that disparate impact analysis applied equally to objective and subjective assessments, the *Watson* court wanted to prevent employers from circumventing the Griggs standard by replacing objective tests with subjective assessments or by combining a subjective assessment with an objective test in the evaluation process.

The employer in the *Watson* case tried to dissuade the court from applying the *Uniform Guidelines* to subjective assessments by arguing that measurement techniques were not sufficiently well developed yet to be applied to subjective assessments. But the court responded that subjective assessments should be subject to statistical scrutiny and Title VII standards, because they could be inappropriately affected by prejudices and stereotypes.

However, the plurality opinion in the *Watson* case gave the impression that the court might apply more rigorous measurement standards to objective tests than to subjective assessments. The *Watson* court stated as follows:

> In the context of subjective or discretionary employment decisions, the employer will often find it easier than in the case of standardized tests to produce evidence of a "manifest relationship to the employment in question." It is self-evident that many jobs, for example those involving managerial responsibilities, require personal qualities that have never been considered amenable to standardized testing. In evaluating claims that discretionary employment practices are insufficiently related to legitimate business purposes, it must be borne in mind that "[c]ourts are generally less competent than employers to restructure business practices, and unless mandated to do so by Congress they should not attempt it." (p. 2791)

Citing another landmark employment case, *Washington v. Davis*, the *Watson* court suggested that employers might not always be required to conduct formal predictive validation studies relating subjective assessments to job performance. In the *Davis* case, the court had allowed an exam given to police academy applicants to be validated by correlation with training program success rather than job success. The *Davis* court justified its position by observing that the connection between training program and job success was obvious.

But the three judges who dissented in the *Watson* case argued forcefully that the standards should be the same for objective tests and subjective assessments:

> Allowing an employer to escape liability simply by articulating vague, inoffensive-sounding subjective criteria would disserve Title VII's goal of eradicating discrimination in employment. It would make no sense to establish a general rule . . . which left the assessment of a list of general, character qualities to the hirer's discretion. Such a rule would encourage employers to abandon attempts to construct [objective tests] for the shelter of vague generalities. (p. 2796-97)

In a prior teacher testing case involving an objective assessment, *United States v. South Carolina*, the court also took a flexible approach. In this case, applicants for teaching certificates challenged the requirement of a passing score on the National Teacher Examination. South Carolina had obtained evidence relating the skills tested to those taught in teacher preparation programs in the state, but had not correlated the test with any measures of successful teaching. Nevertheless, the court upheld the testing requirement primarily on constitutional rather than Title VII grounds.

Shifting Burdens of Proof

Disparate impact challenges under the *Griggs* standard required the employer to demonstrate a compelling reason for using a challenged hiring practice with an adverse impact on a protected group. This requirement was based on a stringent standard that was difficult for the employer to satisfy.

But in a 1989 case, *Wards Cove Packing Co. v. Atonio*, the Supreme Court altered this burden on the employer. The *Wards Cove* court required the employer to meet a less stringent standard of producing evidence to justify its employment practices. Although it was easier for employers to satisfy this burden, employees could still prevail if they could identify an equally effective but less discriminatory alternative.

The 1991 Civil Rights Act

Congress did not like the less stringent standard announced by the *Wards Cove* court. In the 1991 Civil Rights Act, Congress mandated the *Griggs* standard for disparate impact challenges. This legislation placed the full burden back on the employer to convince the court that hiring practices with adverse impact on protected groups were job related for the position in question and consistent with "business necessity." In addition, Congress provided that in cases where the components of a decision-making process are not amenable to separate analysis, the employee would not be required to identify the specific practice causing the discriminatory result. In future cases, the Supreme Court will have to reconcile the conflicting *Griggs* and *Wards Cove* holdings with the language in the Civil Rights Act. However, despite potential disagreements about the standards to be applied in disparate impact challenges, it appears that the federal courts will continue to recognize disparate impact challenges to both objective tests and subjective assessments.

Legal Perspectives from Higher Education Cases

In a salary/promotion case involving subjective evaluations of African-American university employees, a federal court held that subjective evaluations are unlawful only if they are not job-related. In this case, the court found that the challenged subjective assessments by the African-American employee's supervisors were fair and nondiscriminatory. In an earlier case, the court had ruled that Title VII prohibits placing employees at a disadvantage because of their race or gender. If there is evidence of such disadvantage, the burden is on the employer to convince the court that the result would have been the same if the disadvantaged employee had been a Caucasian male.

Dismissal from an Academic Training Program

Several cases have challenged dismissal from a college or university academic program. In such cases, courts generally scrutinize the process to determine procedural fairness, but defer to subjective evaluations of institutional personnel for substantive academic decisions.

For example, in *Schuler v. University of Minnesota*, a Ph.D. student challenged dismissal from a doctoral program for failure to meet academic standards. The student had failed an oral exam that required 4 out of 5 positive faculty votes. The student alleged due process violations for failure to tape record the oral exam to produce a reviewable record and for failure to provide the student with written criteria for evaluation of the oral exam performance.

The court rejected these claims, stating that dismissal from an academic program does not receive full Fourteenth Amendment protection. The only requirement, the court stated, was prior notice of inadequate performance and a deliberate and carefully considered decision to dismiss the student. In this case, the court held that this requirement had been met, that the university was not required to provide evaluation criteria in advance, and that the university could choose whether to use objective or subjective evaluations. Because due process does not require a hearing before academic dismissal, the court further ruled that the university was under no obligation to record the oral exam.

In *Board of Curators v. Horowitz*, a candidate's academic dismissal was challenged on substantive due process grounds. The court rejected the claim, stating that the dismissal was not arbitrary, capricious, motivated by bad faith, or lacking in professional judgment.

In still another academic dismissal case, *Hankins v. Temple University*, an African-American female fellow challenged dismissal from a postgraduate program at Temple University Medical School. The court held that the only process to which she was entitled was an informal faculty evaluation. The court found that meetings with faculty and explanatory letters provided the requisite notice and opportunity to be heard.

Revocation of a University Degree

Revocation of a university degree occurs rarely and typically only after serious allegations of misconduct. For example, the University of Michigan revoked a master's degree from a graduate who was later found to have fabricated the data for the required thesis (*Regents of the University of Michigan v. Ewing*). In holding that the regents of the university had the authority to revoke a degree for cause without a court proceeding, the court stated that otherwise the university would be providing the public with a false certification of the graduate's qualifications. The court indicated that the only process due the graduate was notice of the intended revocation, notice of the university's evidence of serious misconduct, and an opportunity to respond to the allegations. These minimal due process actions were required because the court found that revocation of a degree is a combined academic/ disciplinary action. However, the court refused to find a substantive due process violation, because the professional academic judgments made regarding the graduate's conduct were not arbitrary or capricious.

Other Legal Perspectives on Performance Assessments

Assigning Language Jobs Via Phone Interviews and Ethnicity

In a recent case involving subjective assessments used to assign FBI agents to cases, *Perez v. FBI*, the court held that the FBI's procedures discriminated against Hispanic agents by exploiting their language skills. Hispanic agents who had been hired for skills other than language translation were given short, nonstandardized phone interviews to assess their Spanish language skills. The questions asked during the brief phone interviews varied by administrator and no systematic written evaluation was produced. The Hispanic agents who were assessed to have Spanish language facility were given undesirable assignments. These tasks were undesirable because they were outside of the agents' fields of expertise and did not count toward promotion. Hispanic agents received few promotions and remained in entry-level positions much longer than their non-Hispanic counterparts.

In a class action suit, the Hispanic agents alleged that the phone interviews inaccurately assessed their Spanish language skills, that many Hispanic agents with other specialties were not fluent in Spanish, and that the Hispanic agents were being treated differently from their non-Hispanic counterparts. The suit further alleged that these actions by FBI administrators violated the Title VII ban against discrimination in employment.

The *Perez* case was unique in questioning the validity of an assessment claimed to overrate the skills of historically disadvantaged candidates. Normally, the claim is that the scores of historically disadvantaged group members are too low and are thus preventing them from qualifying for particular jobs. But in the *Perez* case, the court found that the Hispanic agents were being arbitrarily assigned to the least desirable jobs outside the normal promotion track.

The court held that the FBI's subjective phone assessments and subsequent language-based assignments were invalid and discriminatory.

Nonrenewal of Teaching Contracts

Most performance assessments evaluate examinees directly. However, in a few teacher evaluation cases, administrators have proposed indirect performance assessments that evaluate teaching effectiveness by the level of achievement of the teachers' students. Challenges to teacher evaluation systems in Iowa and Missouri provide contrasting views of the appropriateness of using student achievement test score data as a performance assessment measure of teaching effectiveness.

In a 1973 Iowa case, *Scheelhaase v. Woodbury Central Community School District*, a nontenured teacher with ten years of experience was not rehired because her students' test scores for the previous year were too low. The federal court upheld the teacher's dismissal because the school board was not required to show cause when it failed to renew a teaching contract and because its actions were based on the good faith, expert opinion of the superintendent.

The 1987 Missouri case, *St. Louis Teachers Union v. St. Louis Board of Education*, involved tenured teachers who were given unsatisfactory performance evaluations because their students' standardized achievement test scores were too low. The teachers argued that this procedure was unfair because the test and norms were eight years old, the student test had not been validated for evaluating teachers, and the resulting unsatisfactory ratings were arbitrary and capricious. The court ruled against the school system in refusing to dismiss the teachers' case.

Measurement Issues

More than a decade of testimony by expert witnesses has made the courts more knowledgeable about measurement issues and more conversant with its associated technical terminology. Thus, courts are more willing now to scrutinize closely evidence of a test's reliability and validity. However, courts still depend on expert judgment and professional standards to set the boundaries of appropriate and inappropriate practice.

Professional Standards

Different sets of professional standards may be emphasized in litigation, depending on the type of high-stakes assessment being challenged. The EEOC *Uniform Guidelines* will be prominent in challenges to employment assessments. The APA/AERA/NCME *Standards for Testing* will be important in challenges to diploma and licensure tests. The following sections outline the major measurement issues relevant to the development of performance assessments that meet professional standards and are legally defensible.

Testing as a Vehicle for Curricular Reform

It is well documented that statewide diploma tests and the attendant accountability pressures affect what teachers do in their classrooms. Therefore, some educational reformers who want to change teachers' instructional practices significantly have suggested that statewide assessments should be models of good teaching practices. The problem with this position is that a single assessment may not adequately and simultaneously serve the dual goals of individual student evaluation for a diploma and evaluation to improve instruction.

A single assessment used for both purposes (awarding diplomas and modeling preferred instructional practices) confuses the ends and the means. Appropriate instructional practices are the means by which students may achieve important knowledge and skills, but they do not guarantee it.

For example, language arts specialists may believe that the best way to become a good writer is to work collaboratively with other students. Thus, they may want to design assessments that replicate this process. But there may be serious problems with using a collaborative exercise to determine whether or not a student should be awarded a high school diploma. These problems include incongruous levels of measurement and potential unfairness.

When a collaborative exercise is used to make decisions about an individual student, there is a mismatch between the level at which data are collected and the level at which those data are applied. Typically, collaborative efforts receive a single group evaluation. But such group evaluations do not provide appropriate information for determining the level of achievement of a single individual within the group.

Collaborative exercises may provide an erroneous view of the achievement of an individual member of the group, because it is difficult to separate out the contributions of individual members. For example, suppose a poor writer with good social skills appears to work effectively with the group and the group produces an excellent piece of writing by fully utilizing the skills of an individual in the group who is an excellent writer. Or suppose a shy student who is an excellent writer chooses not to participate in the collaborative exercise. Should the former student be awarded a diploma and the latter student denied one? The issue for policymakers is whether satisfactory participation or lack of satisfactory participation in a collaborative process should determine whether a student has adequate writing skills to be awarded a high school diploma.

Because the denial of diplomas to students who have not successfully completed a preferred instructional process is an indirect method for affecting what teachers do in the classroom and because a single instructional process may not be appropriate for all students, policymakers might better achieve their goals by separating the instructional and diploma aspects of assessment. Assessment instruments used to award diplomas should focus on the achievement of skills by individual students. Other classroom evaluation instruments can be implemented to evaluate curricular reform efforts and to provide models of instructional processes that policymakers want teachers to adopt.

Validity

The type of validity evidence necessary to support classroom performance assessments is different from that required for a high-stakes performance assessment used for individual diploma decisions and institutional accountability. In the classroom, decisions are usually low-stakes because they can be changed easily in the face of new or conflicting data. But in high-stakes applications, a single assessment can change an individual's future.

Educational reformers and curriculum specialists generally have been enthusiastic about performance assessment. It certainly has encouraged classroom teachers to expand their instruction beyond rote memorization and repeated drill. However, the classroom appropriateness and advantages of performance assessment can be translated into valid large-scale assessment only at great expense in time and resources.

The rhetoric that seems to suggest that performance assessment is a panacea for past testing problems is not data-based. It is couched in glowing generalities related primarily to instructional objectives. But as statewide programs begin to implement performance assessments, they are finding out that doing it right is a much more complex task than originally anticipated. As data are finally being collected and made public, states are finding that they have not achieved the promised benefits and that the technical properties of the new assessments are far below the quality standards demanded by high-stakes, large-scale testing (Braun, 1992; Harp, 1993). Performance assessments cost more, require more time to develop, are harder to standardize, are very difficult to equate, create content sampling and test security problems, may increase adverse impact on historically disadvantaged groups, and require large expenditures of resources to train raters and develop defensible scoring criteria.

When reformers began lobbying for performance assessment, they stated that the authenticity of performance tasks would make it unnecessary to collect validity and reliability evidence. As long as the performance tasks appeared to tap real-world skills and were tasks that should be taught, the reformers argued that traditional concepts of validity were outdated and unnecessary. The argument seemed to be that "face validity" was enough.

However, face validity is superficial; it means only that the assessment is "valid on its face" or *appears* at first glance to measure the right abilities. But absent evidence of what a performance task actually measures, whether it matches current instruction, and how consistently it is being scored, one cannot know with certainty whether a given set of performance tasks satisfies relevant validity standards. The substitution of performance tasks for traditional multiple-choice items does not by itself guarantee that the resulting assessment will match the goals of a training program, classroom instruction, or relevant job skills.

For example, suppose that students have been taught how to write checks but the performance task involves filling out a portion of a tax form. The form of the performance task is "authentic" because it reflects a skill that is important for adults in the real world. But this performance task is measuring skills that students have not had an opportunity to learn. Analogously, if the performance task is to make change when a simulated customer hands the examinee $20 to pay for items costing $3.99 and $1.15 plus 6% sales tax, the performance task is closely parallel to a real world situation. But if in the classroom students have learned only to make change when $1 is tendered for a single item costing less than $1 with no sales tax, the performance task will not match the instruction and will be unfair. This performance task would also be unfair if given to an applicant for a custodial job that never requires the person to handle money or make change.

Some measurement experts also question whether the tasks labeled as "authentic" performance assessments really reflect the intended types of real world tasks. Such tasks have been variously described as brain teaser exercises, multiple-choice items without distractors, interesting trivia, and window dressing lacking in verisimilitude. Observers have wondered whether the task of "designing a fast food restaurant" reflects the kinds of skills most students will need and whether it can provide a useful aggregate picture of the degree of attainment of educational goals. Similarly, an employment task may be "authentic" but have no relationship to essential job skills.

Even if the reformers are correct in their beliefs about the validity of "authentic" performance assessments, courts may be unwilling to take their word for it. That is, courts typically find simple assertions of validity without supporting evidence to be unconvincing. Particularly when there are adverse effects on protected groups, the courts will require adherence to relevant professional standards. Both the *Uniform Guidelines* and the *Standards for Testing* require the user to demonstrate the validity of any assessments used to make decisions about individuals.

What kind of validity evidence will the court require? For tests used to award diplomas, the court in the *Debra P.* case was very clear about the requirement for curricular validity. Under this standard, test users will be expected to show that all students had the opportunity to learn the tested skills. If performance tasks assess skills not yet in the curriculum, they will not meet this standard, even if they represent skills that educators believe ought to be taught. Only skills actually taught in the state's classrooms can be included on a multiple-choice or performance assessment used to award diplomas.

This legal standard does not preclude the state from adopting new curricula that include performance assessment skills. But before the new skills can be included in a diploma test, they must be communicated clearly to all districts and sufficient notice must be given for their implementation. Changes in tested skills adopted one year for implementation the next year probably will not be deemed to have given students sufficient notice to prepare or to have given districts sufficient time to meet curricular validity standards. Although the court is likely not to require the state to demonstrate that all skills were taught by all teachers in all classrooms to all students, it will expect the state to demonstrate that most teachers recognize the importance of the assessed skills and have included them in their instruction.

Opportunities for retakes and remediation that succeed in decreasing failure rates for historically disadvantaged groups also will be important in judging the fairness of a performance assessment program. And, of course, all performance tasks must have content validity; that is, they must match the objectives that they are intended to measure.

For licensure and employment tests, curricular validity will not be a prime concern. Here, the courts will be looking for either content (licensure) and/or predictive (employment) validity. That is, performance tasks for these types of tests must either contain content minimally necessary for successful job performance or accurately predict those who will be successful and unsuccessful in the job. Again, the courts will not accept a mere assertion that a set of performance tasks is job-related or necessary to protect the public. The state must complete a detailed job analysis or study demonstrating the relationship between job requirements and assessed skills.

Although courts have sometimes accepted evidence relating the test to a training program, this acceptance has occurred when there is little debate about the relationship between the training goals and the necessary job skills. While it is probably most desirable to demonstrate a direct connection between test content and job skills (content validity), predictive validity may be useful for performance assessments when there is disagreement about job performance criteria, the assessed skill is not amenable to direct instruction, or the assessed skill is not sufficiently documented in the literature to provide a construct interpretation.

Therefore, states must collect validity evidence for performance assessments just as they did for traditional multiple-choice exams. The alleged "authenticity" of performance tasks cannot substitute for the validity evidence that courts have come to expect and demand. Particularly for diploma and licensure tests, the court will balance protection of the public under the state's police powers against the degree of infringement on individual rights and potential unfairness to protected groups.

Because measurement professionals may continue to disagree on exactly what validity evidence must be collected, the courts may continue to balance the burden on the state to produce appropriate validity evidence against the burden on the challenger to produce evidence of invalidity. Both the *Watson* and *South Carolina* cases discussed in earlier chapters demonstrate that the courts may accept good faith efforts to apply state-of-the-art technology in cases in which validity evidence falls short of the ideal. But states should remember that a good faith effort means knowing the standards and satisfying them in spirit, as well as all particulars that are reasonably attainable.

Content Sampling

The reliability or consistency of a performance assessment has two equally important components: rater consistency/accuracy and content domain sampling. Content domain sampling refers to the selection of a sample of specific content to represent the full domain of knowledge and skills to which the test user wishes to generalize.

Time and resource limitations prohibit the test user from assessing all of the knowledge and skills contained in an educational program or required for an occupation. But the more educational goals or job skills assessed, the more accurate will be the judgment of the examinee's abilities relative to the entire domain.

However, given a fixed amount of testing time, performance assessments are able to sample an even smaller fraction of the domain of desired behaviors than multiple-choice tests. For example, in a one-hour period, one might be able to ask only 2 or 3 short essay questions covering 2 or 3 different topics. But in that same time period, one might be able to ask 50 multiple-choice questions covering 50 different topics.

The trade-off in content sampling between multiple-choice items and performance assessment tasks is a difference between depth and breadth. The performance tasks provide a comprehensive view of understanding in a few areas, whereas multiple-choice items provide snapshots of understanding for a much larger set of domain topics. For some subjects, such as writing, the skill being assessed requires the depth and realism of a performance task. But for other subjects the writing or demonstration aspects of a performance task may require testing time that is disproportionate to the amount of information gained. When a test user is considering substituting performance tasks for multiple-choice items, it is important to ask whether the gain in depth is worth the loss in breadth.

The number of unique concepts that can be sampled on a test is important because it affects the magnitude of the errors of measurement. Measurement error occurs when extraneous factors cause a test score to be lower or higher than the student's actual ability. Longer tests have smaller measurement errors than shorter tests. Because performance tasks usually require the examinee to construct rather than recognize a correct answer, performance tasks require greater response time. Therefore, fewer performance tasks can be administered in a given amount of time. This limitation results in fewer samples of behavior from which to generalize to the domain of interest and a greater opportunity for inappropriate teaching to the test.

To illustrate this point, consider the following hypothetical situation. Suppose a domain of interest contains 100 unique concepts and the test user has both short essay and multiple-choice items available for test construction. Assume further that with the reading, organizing, and writing involved, each essay question would require approximately ten minutes to answer but that, on average, only one minute is required to read and answer each multiple-choice item. Given these assumptions, a one-hour essay exam could sample 6 concepts, while a one-hour multiple-choice exam could sample 60.[1]

Now suppose an examinee understands 90 of the 100 concepts but has not mastered the remaining 10. If the short essay and multiple-choice exams each happen to sample 2 of the 10 concepts that the examinee does not know, the examinee will be much more seriously penalized on the essay exam than on the multiple-choice exam. Assuming that the examinee correctly answers all items testing the known concepts and all items are weighted equally, the percent correct scores on the essay and multiple-choice exams will be 67% and 97%. Clearly, the multiple-choice exam would provide a more accurate estimate of the examinee's actual knowledge of 90% of the domain.

---

[1] Although longer, complex essay questions might cover 2-3 concepts, for purposes of this example the author assumes that each ten-minute essay covers a single concept.

As the above example illustrates, when tests contain few items, individual student weaknesses may be more easily hidden if the sampled concepts happen to be strengths or may be disproportionately emphasized if the sampled concepts happen to be weaknesses for that student. Put another way, different samples of 60 items from the same domain will yield more similar estimates of total knowledge than will different samples of six essays. Theoretically, all six essays could come from the 90 known concepts or from the ten unknown concepts, so the examinee's essay score could range from 0% to 100% (assuming no partial credit). But a maximum of only ten multiple-choice items could come from unknown topics, so the possible range of multiple-choice scores would be 83% to 100%.

This problem is similar to a situation in which one wants to know how 100 people feel about an issue. Asking 60 is more accurate than asking only 6. The point is that assessment is less accurate when there are fewer measurements.

Thus, under certain circumstances, depending on the definition of the domain in which achievement is being assessed, the essay exam is much more likely to underestimate or overestimate the student's true knowledge of the domain and result in an erroneous decision about the student's competence. In a classroom, other assignments, tests, or projects would be available for evaluating the student. But in a large-scale testing program, if achievement is underestimated the student's only redemption may be to retake the test and hope for a more favorable sampling of content. And even when the content sampling of performance tasks is adequate, research has suggested that high scores on specific performance tasks may not correlate highly with other performance tasks or generalize to other content within the same domain.

In the previous hypothetical example, the narrower range of possible multiple-choice scores around the student's true 90% ability reflected the greater consistency and higher content sampling reliability of a longer test. This example suggests that performance tasks should be reserved for those concepts for which the performance skill is an essential component in demonstrating the requisite knowledge.

The content sampling issue also has relevance to critics' concerns that testing narrows the curriculum. These concerns arise because of the tendency of some teachers under accountability pressures to concentrate their instruction on the specific concepts tested. For example, using the previous illustration, if the statewide curriculum comprises 100 concepts, but only the same 60 concepts are known to be tested each year, some teachers may not teach the other 40. But if the multiple-choice test is replaced by a performance assessment (the six essay questions), the set of concepts on which a teacher may inappropriately focus may be considerably smaller.

One way to address this problem is to select a different sample of the 100 concepts to test each year. If a teacher did not know which concepts would be sampled, the teacher would be more likely to teach all 100. But teachers tend to fight such proposals, because they feel that they are entitled to know ahead of time exactly what will be tested. In addition, changing both the items and the concepts each year requires additional technical work that may significantly increase the cost of the assessment.

Finally, for some performances, such as laboratory experiments, a small number of different performance tasks may be available for inclusion on different test forms. On the other hand, an almost infinite number of possible multiple-choice items could be used to generate new test forms. If performance assessment is selected, the performance tasks will have to be repeated often in subsequent forms of the assessment. With multiple testings each year, the entire item pool soon may become known and students may be drilled over and over on the small set of experiments that might be on the assessment. It then becomes impossible to tell whether the students have the skills that the performance tasks were designed to assess or whether they are repeating what they have been told repeatedly in class.

Scorer Reliability

In addition to content sampling, the other component of reliability that must be addressed in performance assessment is scorer reliability. The potential for scorer unreliability in performance assessment is much greater than in multiple-choice testing, because performance assessment scores depend heavily on fallible human judgment.

For a multiple-choice test, content experts agree in advance on the correct answer to each item. When examinees' tests are scored, the same key is used for everyone, so potential scoring errors are clerical. Most clerical errors can be eliminated with machine scoring.

But in performance assessment, the task responses are rated using human judgment. Because humans are fallible, so are the scores they assign. Thus, for performance task ratings, one must consider whether an examinee might receive a significantly different score if the performance task were rated by a different scorer.

The ways in which scorers can behave unreliably are varied. They may impose different standards for the same score. One scorer may be very lenient while another is very tough. Some scorers may be willing to use all of the score points, giving some very high scores and some very low scores. But other scorers may prefer not to use the extremes and may assign scores that are close to the mean or central score.

Scorers also may introduce unreliability into the scoring process as a result of the context in which the performance is rated. After rating three good performances, a scorer may be more inclined to give an average paper a lower score than it deserves. Conversely, after rating three poor performances, a scorer may give an average paper a higher score than it deserves. Scorers also may be affected by fatigue. The first performance a scorer evaluates may be scored differently than the hundredth performance. Or a scorer may start off rigid and become more flexible. Consequently, without safeguards, papers scored earlier may receive lower grades.

Scorers may be directly or indirectly influenced by factors that are not part of what the test user intends to measure. Such factors include those that operate when the rater knows or can see the examinee (e.g., appearance, sex, race, sexual preference, prior reputation, shyness, assertiveness, grooming, physical appearance) and those that can operate any time (e.g., handwriting, spelling, nonstandard English, context effects). When the performance is written, the former list of factors can be controlled by anonymous scoring, but the latter list of potential "biases" remains problematic.

In some cases, these characteristics may be a legitimate part of the skill being assessed. For example, correct spelling may be important in an English composition. But individual characteristics such as physical appearance typically have no place in an academic rating.

Even those scorers who know better can subconsciously alter their views of a performance according to their own biases and prejudices. The debates over the past several decades regarding affirmative action, sexual harassment, and homosexuals in the military demonstrate unequivocally that people have strong and opposing views about certain characteristics. As a result of characteristics offensive to those in power who subjectively evaluate performance on the job, people have been refused employment or have lost their jobs because they were from the wrong race/ethnic group, refused sexual advances, or were openly homosexual. It is common for those with better handwriting to be perceived as better writers or those with no accent to be perceived as better news broadcasters. Indeed, it is difficult to focus on the substance of a response if that response is difficult to read or understand.

However, scorers can be trained to apply uniform standards and to minimize potential biases. The training must be intensive and thorough and scorers must be periodically rechecked to ensure that they are maintaining the same standards from the first performances scored to the last.

Before training begins, one must select scorers who have appropriate content qualifications to rate the performances. For example, a diving coach would not be asked to rate a musical recital. Raters also must be independent; they must not have a vested interest in the outcome of the scoring process. In high-stakes testing this means, for example, that teachers should not score the performances of their own students.

After the requisite number of qualified scorers has been selected, the training usually includes three phases: explanation of the standards, practice, and qualifying. The standards to be applied to each performance must include specific criteria, samples of performances at each score point, and guidelines for weighting the components of a performance. Scorers who successfully complete the three-phase training are randomly assigned performances to rate.

Although a single, highly reliable rating for each performance may be satisfactory in some contexts, when feasible at least two scorers should rate each performance. Trainers should be available during scoring to answer questions and resolve discrepancies. Typically, a third scorer rates a performance for which the two initial scores differ by more than one point. The use of a single scorer for each performance may be appropriate when the additional reliability of a second scorer is not cost effective, the examinee's total score is based on many performances, the performances have a low weight, there is high agreement among raters, or there is provision for a second rater for cases near the passing standard.

When more than one score is obtained for each performance, consistency of scoring can be estimated by calculating the percent of exact agreement and percent of disagreement greater than one point for each scorer. Periodically, or daily for high-stakes performance tasks rated by a single scorer, unidentified verification papers can be included in a scorer's assigned work to check for potential drift from the standards. Scorers who do not maintain standards should be retrained and required to requalify.

In addition to the costs of selecting, training, and rechecking scorer accuracy, the scoring of performance assessments also incurs costs for scorer compensation and housing, collecting, packaging, and distributing task responses to multiple scorers, and entry of scoring results into a data base from which score reports can be produced. The goal of this costly process is to ensure fair and reproducible scores. Any attempts to skimp on any of these items to save costs may result in less defensible scores.

With many departments of education receiving budget cuts because of statewide deficits, a cost/benefit analysis may be helpful prior to embarking on a large-scale performance assessment program. By careful consideration of the assessment areas in which the information gained is worth the substantial costs, limited funds for performance assessments can be carefully allocated to be most cost-effective.

For those skills targeted for performance assessment, careful consideration must be given to the trade-offs in reliability and validity involved in choosing the number of performances to be rated and the number of scorers to rate each performance. Reliability and validity both increase with more performances and more raters. But this relationship remains true only with adequate scorer training and strict adherence to rating criteria. On the other hand, a single performance rated by a single scorer is probably too unreliable to be used for a high-stakes decision such as awarding a diploma. Although cutting corners in performance assessment may achieve short-term cost containment, in the long run the cost of litigation over flawed procedures may erase any savings or place the entire assessment program in financial jeopardy.

In addition to using appropriate scoring procedures, high-stakes test users must document the process in detail and keep accurate records. The purpose of detailed documentation is to increase uniformity by clearly communicating policies and procedures. The goal is to maximize consistency of measurement across tasks and to obtain evidence of a high degree of agreement among raters.

By extensively training raters and developing detailed scoring rubrics, statewide writing assessments have been able to achieve relatively high scorer agreement. To achieve similar results, other subjects will have to follow the same expensive and tedious procedures. When moving from written responses to oral responses or demonstrations, test users also must be vigilant for extraneous factors that might negatively affect the reliability and validity of the scoring process.

In summary, performance test users have multiple responsibilities to (1) document scorer selection and training, (2) document scoring procedures, (3) obtain evidence of scorer agreement, (4) demonstrate that the sampled tasks consistently measure the knowledge and skills being assessed, and (4) establish that the tasks adequately sample the content domain.

Standardization

The purpose of high-stakes testing is to compare performances using common scoring criteria. Comparisons can be made only if all examinees respond to the tasks under standard testing conditions. Standardized written directions for test administration, common sample exercises, identical items/tasks, and maximum time limits are customary elements of a standardized assessment.

If the performance task requires specialized equipment such as lab materials or calculators, fairness also may dictate that all examinees be given adequate opportunity to familiarize themselves with the equipment and its operation prior to the assessment. All examinees also must have equal access to quality equipment. For example, it would not be fair if some students had simple four-function calculators while others had expensive calculators with memories and statistical functions.

Test Security

The rewards and sanctions associated with high-stakes testing may cause some individuals to engage in inappropriate test preparation activities. For example, when a teacher teaches the content of the specific items on the test rather than the full set of knowledge and skills in the sampled domain, the teacher has inappropriately prepared students for the test. Teaching the small set of skills that happen to be sampled from the domain is a much easier task than teaching all of the skills in the domain. As a result, the performance on the sampled tasks is no longer a reasonable estimate of overall domain performance. Over time, such practices result in increased test scores without a concomitant increase in the inferred overall achievement. Put more simply, test scores increase but students have learned less. A few years ago, a national survey discovered the "Lake Wobegon effect" when all 50 states claimed that their students were above average. This phenomenon is discussed in greater detail in Chapter 2.

Inappropriate test preparation activities are not a function of the format of the test items. Accountability pressures can cause both multiple-choice items and performance tasks to be targeted. As indicated in the section on content sampling, the number of potential performance tasks may be finite and resource limitations may dictate that tasks be recycled frequently, which increases the likelihood of inappropriate preparation activities. With fewer performance tasks in circulation and greater overlap from form to form, it is much easier for teachers to remember and teach the content of specific tasks and much easier for unauthorized information on specific content to be passed on to applicants for licensure or certification. This increased likelihood of inappropriate preparation provides an advantage to those who receive special preparation and invalidates the interpretation of their test scores.

Some advocates of performance assessment argue that teachers who teach performance assessments are teaching the right skills in a more desirable way. They argue that teaching an experiment or solving a problem is better teaching than the rote and repetitive drills some teachers use to teach the content of multiple-choice items. But no matter what the format of the items, teaching the sample is not the same as teaching the domain, and teaching the content of performance tasks from an assessment is still dishonest. For example, if students have repeated a science experiment three times in the two weeks before they are subjected to a performance assessment, one can infer little about their ability to design an appropriate experiment, make predictions about results, or interpret findings. One can generalize only to the set of tasks on the assessment and one can not generalize to the larger domain from which the tasks were sampled.

Appropriate test preparation activities should follow the letter and spirit of ethical professional standards and should increase test scores only when student achievement of the content domain has increased. A high-stakes statewide testing program can take several steps to discourage and minimize the effects of inappropriate test preparation activities and outright cheating. However, such measures are costly and time-consuming and may delay the reporting of test results. These steps are listed in point 14 of the Chapter 2 recommendations.

Potential Bias

Traditional multiple-choice tests have one potential source of bias (the items), whereas performance assessments have two (task content and scorer prejudices). Hence, assessment programs that include performance assessments must work extra hard to ensure accuracy and fairness.

Differential performance between majority group members and historically disadvantaged group members has been a serious area of concern in traditional testing programs and will continue to be an issue for performance assessment programs. Through efforts to eliminate offensive items and vestiges of segregation, the gap between majority and historically disadvantaged group test performance has narrowed on traditional multiple-choice tests but may widen again with the introduction of performance assessments.

Several well-researched, professionally accepted technical methods exist for detecting such differential performance in individual multiple-choice items. Differential performance measures similar to those used for multiple-choice items may be adapted for use in performance assessment contexts. As with multiple-choice items, performance task measures of differential performance must avoid simplistic notions such as the discredited *Golden Rule* procedure. Appropriate procedures for identifying differential task performance must hold ability constant while measuring group performance differences.

The *Golden Rule* procedure called for eliminating items that were difficult for African-Americans and preferring items with total African-American/Caucasian percentage correct differences that are less than 15%. This type of rule threatens the content validity of an exam by potentially eliminating all tasks in a difficult or more complex subskill, particularly when the set of available tasks is relatively small.

Unfortunately, even the most technically correct procedures for identifying differential performance by subgroups may not produce identical results on the same set of tasks. Although there may be overlap, each method of detecting differential task performance will identify a different set of potentially biased tasks. Thus, these statistics should be interpreted only as indicators for further scrutiny of the identified tasks. Generally, a group of content experts with substantial representation from historically disadvantaged groups gives further scrutiny to items identified as biased by statistical procedures. If the representatives from historically disadvantaged groups believe that an identified task is appropriate and cannot explain the performance differences, then the task can be retained.

In most cases, items with large discrepancies between historically disadvantaged group and majority performance are also "flagged" as "bad" items by traditional item selection procedures. But occasionally statistical procedures can erroneously identify an item as biased, which may explain why 60% of 30 = ___ may be identified as a biased item when 70% of 20 = ___ is not. It would seem that only the tasks with the most extreme majority/historically disadvantaged group differences (the outliers) should be "flagged" for further scrutiny.

Like multiple-choice items, performance tasks also may contain language or concepts for which a particular group has a cultural disadvantage. But in addition to potential task characteristic biases, the response formats and scoring of performance assessment tasks may create an additional disadvantage for some historically disadvantaged groups. For example, persons with limited English skills, heavy accents, tics from nervous disorders, or disfigurements may be at a disadvantage when oral responses are required. Poor writers may have difficulty with lengthy written responses. Shy persons may hold back and not participate fully in a cooperative experiment. If scorers have not been thoroughly trained and provided with detailed standards, personal prejudices may cause some responses to be rated lower than they should be. Any number of irrelevant examinee characteristics that become obvious to scorers through the performance response may inadvertently affect scores. These potential biases may be difficult to detect and are probably impossible to eliminate completely.

Other Technical Issues

High-stakes testing programs need alternate forms to maintain test security. For alternate forms to be fair to persons tested at different times, they must be equated to a common scale. Equating adjusts for minor differences in difficulty between test forms that measure the same content.

Procedures for equating multiple-choice tests are well researched and technically defensible. But they are often difficult to explain to legislators and the public, who tend to distrust such manipulations. Rather than understanding that the passing standard is being maintained at the same level for all examinees, the public may believe that the state is using equating to raise or lower the passing standard artificially so that more or fewer examinees will pass.

Except for writing, methods for equating performance tasks are not well researched. Such methods tend to be much more complex and to have larger equating errors. This situation may in part be a function of the greater unreliability of the performance responses on which such methods are based. Until equating technology can be perfected for performance assessments, large-scale programs will have difficulty ensuring the fairness of multiple test forms. Moreover, in addition to adjusting for differences in task difficulty, adjustments for differences in the leniency or stringency of scorers, particularly when rating criteria are very general, may be needed.

Summary

Prior litigation in related areas suggests that courts will apply the *Uniform Guidelines* and *Standards For Testing* when performance assessments are challenged. Although courts may be a bit more flexible in their expectations for performance assessments, states would be well-advised to proceed cautiously and to implement only those new assessments for which adequate technical data are available. This precaution is particularly critical if disparate impact on historically disadvantaged groups is substantial or increases when a new performance assessment is implemented.

Adequate due process notice and appropriate validity evidence will continue to be required for performance assessments. The courts also may require evidence of scorer reliability under the fundamental fairness standard.

The subjective assessments that the courts have invalidated in the past have involved egregious procedural violations. Rarely has a court addressed the substantive and technical adequacy of a subjective procedure. However, because of the availability of professional standards and experts willing to testify about any flaws in a testing program, it is unlikely that a court would assume the validity of a performance assessment without appropriate evidence that comports with the court's notion of "common sense." States should be cautious about making unsubstantiated claims about the advantages of performance assessments. But states should not be totally precluded from making good faith attempts to advance the state of the art.

Recommendations for Developing and Implementing Legally Defensible Performance Assessments

Designing appropriate data collection strategies during the developmental phases of an assessment program is often much easier than trying to collect the required data after a lawsuit has been filed. Knowing what is likely to be challenged and being prepared for such challenges can facilitate settlement and dissuade challengers from initiating protracted court battles. Following professionally accepted standards and carefully documenting all procedures demonstrate good faith. The following are general recommendations for increasing the legal defensibility of performance assessment programs:

(1) Follow the recommendations for diploma testing given in Chapter 2, including the test security guidelines listed under point 14.

(2) Follow the recommendations for addressing differential item performance given in Chapter 3.

(3) Follow the recommendations for developing testing accommodation policies given in Chapter 4.

(4) Provide advance notice of assessment formats and criteria for evaluating performances.

(5) Implement only those assessment procedures for which adequate data are available to document that professional standards have been met. Follow the advice of the technical advisory committee at all stages of the process of changing curricula and tests.

(6) Conduct a cost-benefit analysis to determine the areas in which performance assessment will provide important, unique information at affordable cost.

(7) Consider the potential adverse impact on historically disadvantaged groups and develop strategies for addressing the problem. Pay particular attention to potential scoring "biases" due to personal appearance, race, gender, accents, nonstandard English, poor handwriting, and so on. Use anonymous scoring whenever possible.

(8) Document opportunity to learn or job relatedness before using assessment scores to make high-stakes decisions.

(9) Administer performance tasks under standardized conditions to ensure fairness to all examinees.

(10) In addition to the usual content and bias reviews, carefully consider potential confounding of the task performance due to language deficiencies, writing or speaking

deficits, personal and cultural reactions to the task, knowledge and familiarity with equipment, and other situational variables.

(11) Obtain consensus among content experts for detailed scoring criteria and train scorers to apply the criteria consistently and accurately.

(12) When feasible, obtain at least two scores for each performance and develop a procedure for identifying and resolving scorer discrepancies. When only one score is obtained for each performance, that score should be highly reliable, the total score should include multiple performances, the performances should have low weight, and/or a second score should be obtained for performances near the passing standard.

(13) Periodically and systematically recheck the ratings of each scorer for consistency and accuracy.

(14) Employ sufficient numbers of tasks and raters to ensure adequate content sampling and reliable scoring.

(15) Plan in advance for the scheduling of assessment development activities, necessary data collection, scorer training, and other contingencies so that adequate fiscal and human resources can be appropriated.

References

See also references from earlier chapters that are not repeated here.

Cases and Statutes

Board of Curators v. Horowitz, 435 U.S. 78 (1978).

Civil Rights Act of 1991, Pub. L. No. 102-166, 42 U.S.C. § 2000e.

Crook v. Baker, 813 F.2d 88 (6th Cir. 1987).

Crump, et al. v. Gilmer Independent School District, 797 F. Supp. 552 (E.D. Tex. 1992).

Dalton v. ETS, 588 N.Y.S.2d 741 (Sup. Ct. 1992).

Edgewood Independent School Dist. v. Paiz, 856 S.W.2d 269 (Tex. App. 1993).

Fisher v. Procter & Gamble Mfg. Co., 613 F.2d 527 (5th Cir. 1980).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Hankins v. Temple Univ., 829 F.2d 437 (3rd Cir. 1987).

Helbig v. City of New York, 597 N.Y.S. 2d 585 (Sup. 1993).

Perez v. FBI, 707 F. Supp. 891 (W.D. Tex. 1988).

Regents of the Univ. of Mich. v. Ewing, 474 U.S. 214 (1985).

Rowe v. General Motors, 457 F.2d 348 (5th Cir. 1972).

St. Louis Teachers Union v. St. Louis Bd. of Educ., 652 F. Supp. 425 (E.D. Mo. 1987).

Scheelhaase v. Woodbury Central Community School Dist., 488 F.2d 237 (1973).

Schuler v. University of Minn., 788 F.2d 510 (8th Cir. 1986).

United States v. South Carolina, 445 F. Supp. 1094 (D.S.C. 1977), aff'd mem, 434 U.S. 1026 (1978).

Wards Cove Packing Co. v. Atonio, 109 S. Ct. 2115 (1989).

Washington v. Davis, 426 U.S. 229 (1976).

Watson v. Ft. Worth Bank & Trust, 108 S. Ct. 2777 (1988).

Williams v. Austin Independent School Dist., 796 F. Supp. 251 (W.D. Tex. 1992).

Articles and Other Resources

Beck (in press). "Authentic Assessment" for Large-Scale Accountability Purposes: Balancing the Rhetoric. *Educational Measurement: Issues and Practice.*

Braun (January 1993). Education Department Under Probe on Possible Skill Test Coverup. *The Star Ledger*, *22*, 1.

Cooley (1991). State-Wide Student Assessment. *Educational Measurement: Issues and Practice*, *10*(4), 3.

Forsyth (1976). Describing What Johnny Can Do. *Iowa Testing Programs Occasional Papers*.

Frechtling (1991). Performance Assessment: Moonstruck or the Real Thing? *Educational Measurement: Issues and Practice*, *10*(4), 23.

Harp (1993). Widely Mixed Test Results Leave Some in Kentucky Puzzled. *Education Week*, *13*(3), 15.

Mehrens (1992). Using Performance Assessment for Accountability Purposes, *Educational Measurement: Issues and Practice*, *11*(1), 3.

Mehrens, Phillips, & Schram (in press). Statewide Test Security Practices, *Educational Measurement: Issues and Practice*.

Meredith (1984). Writing Assessment in South Carolina: Past and Present, *Educational Measurement: Issues and Practice*, *3*(1), 19.

Phillips (in press). Legal Issues in Performance Assessment. *Education Law Reporter*.

Popham (June 1991). *Circumventing the High Costs of Authentic Assessment*. Paper presented at the annual Education Commission of the States Assessment Conference, Breckenridge, CO.

Popham (1991). Appropriateness of Teachers' Test-Preparation Practices. *Educational Measurement: Issues and Practice*, *10*(4), 12.

Rebell (1990). Legal Aspects of Subjective Assessments of Teacher Competency. In National Evaluation Systems, *The Assessment of Teaching*.

Sachse (1984). Writing Assessment in Texas: Practices and Problems, *Educational Measurement: Issues and Practice*, *3*(1) 21.

Tatel (1992). Civil Rights Act of 1991. *Nolpe Notes*, *27*(2), 1.

Wolf, Bixby, Glenn, & Gardner (1991). To Use Their Minds Well: Investigating New Forms of Student Assessment. In G. GRANT (Ed.), *Review of Research in Education*, 31.

Chapter 6

Anticipated Future Legal Challenges

The United States is a litigious society and educational assessment is no exception. Increasingly, statewide assessment programs are facing legal challenges from a variety of individuals and special interests.

Transition to Performance Assessments

Testing to award diplomas continues to be popular with legislators and the business community. However, the emphasis appears to be shifting from traditional multiple-choice items to performance tasks and process-oriented assessments. If the gap between majority performance and that of historically disadvantaged groups widens as a result of the shift to performance assessment, statewide assessments for awarding diplomas may face a new round of legal challenges.

Although differential performance by itself is not sufficient to invalidate an assessment program, defending such a program against a legal challenge based on alleged discrimination can be costly, time-consuming, and detrimental to public relations. It also can exacerbate test security concerns as challengers seek access to the disputed assessment tasks.

In the past, courts have invalidated subjective assessments based on egregious procedural problems. In general, courts have not applied rigorous measurement standards to performance assessments. Although the *Watson* case plurality opinion suggested that the standards for performance assessment might be somewhat less stringent than for traditional multiple-choice tests, recent federal court decisions have indicated a willingness to apply professional standards to performance assessments. For example, in the *Perez* case, where Hispanic FBI agents were forced to accept less desirable, language-related assignments not valued in the promotion process, the court held that phone interviews used to assess Spanish language skill were invalid on technical grounds. This decision suggests that performance assessments with adverse impact will receive rigorous scrutiny and that claims of authenticity (e.g., the test is valid on its face) will not be sufficient.

Content Challenges

Statewide assessment programs also seem to be facing increasing numbers of legal challenges related to content objections. Such challenges typically allege that test items call for value judgments or promote activities contrary to the students' religious beliefs. For example, in one state, parents objected to a reading passage that described a grandfather sitting on the porch smoking a pipe. The parents argued that the test was promoting smoking. In another state case that is pending, parents have objected to a graph interpretation item where the graph depicted percentages by religious affiliation. In this case, the parents have argued that this item and other social studies test items impermissibly require students to make value judgments contrary to their religious beliefs.

Both of the examples given above illustrate an increasing sensitivity among some parents to perceived "hidden agendas" in statewide tests. These parents believe that the test content is intended to influence the beliefs of their children and that such influence is inappropriate. Although such challenges have generally been unsuccessful in obtaining injunctions against further administration or scoring of the challenged tests, they have caused substantial human resources to be diverted to defending the challenged tests. As diversity increases and special interest groups become more organized and vocal, it may become increasingly difficult for statewide assessment programs to identify noncontroversial content for inclusion in assessment instruments.

State vs. Local Control

Another area of concern affecting statewide testing is the tension between state control and local control. While the state has interests in ensuring meaningful levels of achievement for students receiving diplomas in the state, it can be argued that local educators who actually know a student are in a better position to assess the student's knowledge and skills. There will never be a fail-safe system that will prevent the unscrupulous behavior of some test administrators; this point is illustrated by a recent case in which a principal intentionally altered standardized test scores to prevent students from receiving special education or remedial services (*Helbig v. City of New York*). Thus, in balancing the state's interest and the integrity of local decision-making, one might inquire whether the state's interests are best served by standardized test scores as a sole criterion. This issue may become intertwined with funding challenges faced recently by many state legislatures. Statewide assessment results may be used as evidence in a legal challenge to the fairness and adequacy of public educational funding, or may become a pivotal negotiating point as states seek greater accountability leverage over monies disbursed to local educational systems.

Transfer Students

A collateral issue in statewide testing for diplomas and endorsements concerns whether or to what extent states should honor decisions of other states to deny a diploma or endorsement based on failure to pass all or part of a statewide assessment. It is not uncommon for students failing to pass a statewide assessment to move to another state that does not have such an examination, establish residency, and ask for a diploma. If the student has satisfied local and state requirements for a diploma in the new locality, it may be cheaper and easier to grant a diploma than it is to enroll the student and provide duplicative educational services. Given the legal requirement for curricular validity, transfer students—particularly late in the high school program—can pose difficult dilemmas for states with different testing requirements than those to which the student was originally subjected.

Participation in Graduation Ceremonies

A new wrinkle in the diploma denial debate involves challenges to a local district's exclusion of a student from graduation ceremonies when all graduation requirements have been met except passing the mandatory statewide test. Two recent federal court cases in different districts in Texas reached opposite conclusions on this issue. One federal judge, in *Williams v. Austin Independent School District*, found that the statewide test met the *Debra P.* due process notice and curricular validity requirements and refused to require the local district to allow students who failed the test to participate in graduation ceremonies. This judge specifically stated that participation in a graduation ceremony is not a constitutionally protected property right.

The federal judge in *Crump v. Gilmer Independent School District*, on the other hand, found that the school district might not be able to meet its burden of demonstrating adequate notice and curricular validity. Hence, this latter judge ordered the local district to allow the challenging students who had satisfied all graduation requirements except the test to participate in graduation ceremonies. However, this order was conditional on the district's right to announce at the ceremony which students failed to pass the state test and the district's right to withhold these students' diplomas pending satisfaction of the testing requirement. A subsequent state appeals court case on the same issue, *Edgewood Independent School District v. Paiz*, concluded that decisions regarding which students may participate in graduation ceremonies should be left to local officials.

Accommodations for Persons with Disabilities

The greatest potential for future litigation is probably in the area of testing accommodations for persons with disabilities. In general, there is an inverse relationship between the extent of the accommodations and the validity of the assessment program. But denying accommodations to the disabled runs counter to the public policy goal of including the disabled to the maximum extent possible.

Traditional accommodations for physical disabilities are relatively straightforward and typically not controversial. The area ripe for litigation is the line-drawing process involved in determining whether and to what extent cognitive disabilities should be accommodated. Assessment programs that deny requested accommodations such as a reader for a reading comprehension test administered to a learning disabled student may be challenged in court under the Americans with Disabilities Act (ADA). Although some legal scholars believe that the ADA will be interpreted consistent with prior case law, many advocates for the disabled believe that the ADA has extended the rights of the disabled.

In making decisions about which accommodations to allow the cognitively disabled, policymakers will have a difficult challenge in balancing test validity against the rights of the disabled. Guidance from the courts may be a welcome clarification in this area and may lead some programs to seek judicial clarification of legal standards. In the interim, the threat of legal action may pressure some programs to grant accommodations that measurement specialists believe are invalid.

Another area of controversy in the testing accommodations arena is the degree to which disabled persons are entitled to privacy regarding testing accommodations. Some decision-makers argue for "flagging" any score obtained under nonstandard conditions while advocates for the disabled argue that all "flagging" of scores leads to impermissible discrimination against the disabled. Policymakers confronting this issue will have to balance social concerns regarding the potential for inappropriate discrimination against the potential for the accommodated scores to be misleading to test users. Again, it may be necessary for the courts to clarify the requirements of the ADA.

Conclusion

Because of the anxiety and controversy surrounding most accountability programs, policymakers can expect legal challenges to high-stakes assessment programs. Decision-makers who are informed about the legal and measurement issues surrounding their assessment programs should be able to make legally defensible policy judgments and be adequately prepared in the event of a formal legal challenge.

NCREL

North Central Regional Educational Laboratory
1900 Spring Road, Suite 300
Oak Brook, IL 60521-1480
(708) 571-4700
Fax (708) 571-4716

BEST COPY AVAILABLE